Ludwig-Maximilians-Universität München

Fakultät für Mathematik, Informatik und Statistik

Master Thesis

Modeling and forecasting yield curves

A comparison of published methods

Author:
Selina Reinicke

Supervisor:
Dr. Fabian Scheipl
Department of Statistics
Ludwig-Maximilians-Universität München

München, April 16, 2019

Contents

List of Figures

Abbreviations

ACF	autocorrelation function
ACVF	autocovariance function
AIC	Akaike information criterion
AR	auto regressive
BB	Bundesbank
DL	Diebold and Li
FDA	Functional Data Analysis
FPCA	Functional Principal Component Analysis
GP	Gaussian Processes
i.i.d.	independent and identically distributed
MA	moving average
OLS	ordinary least squares
RMSE	root mean squared error
US	United States (of America)
VAR	vector auto regressive
WN	white noise

1 Introduction

Yield curves display the term structure of yields of a certain class of assets. The term structure of a yield curve describes the relationship between the yields of bonds and their respective term to maturity. In this thesis yield curves refer to zero-coupon governent bonds. These yield curves are highly relevant for the valuation of fixed income products and their evolution delivers important information about the economic situation. Yield curves have both a cross-sectional dimension across different maturities as well as a temporal dimension evolving dynamically over time. (Diebold and Rudebusch, 2013).

Modeling yield curves is very valuable for providing information about market conditions. Formerly proposed approaches could be categorized as either no-arbitrage models or equilibrium models. Diebold and Li (2006) upended those traditions and introduced a factor model focusing on forecasting yield curves. Due to its forecasting performance, it developed to be a benchmark to test new methods against or to develop extensions to account for observed constraints.

The data of yield curves can also be viewed as functional data opening up to further methods of modeling and forecasting. In this thesis, next to the model by Diebold and Li for forecasting yield curves, two approaches applicable generally to functional time series are analyzed. These approaches are based on functional principal component analysis and Gaussian processes, respectively. The different approaches are compared with regards to their forecasting performance applying a measurement of prediction accuracy.

The thesis is structured as follows: in chapter 2 the construct of yield curves is introduced. Chapter 3 presents the three methods and models and provides details on their forecasting framework. The data that was used for comparing forecasting performances is presented in chapter 4. Chapter 5 describes the chosen study design and evaluates the forecast accuracy of the different models. In Chapter 6 the results of this thesis are discussed.

2 Economic background and stylized facts about yield curves

For the analyis of fixed income instruments, obtaining interest rate point forecasts is crucial as it is an essential input for financial modeling and risk modeling. Analysts and traders base their conclusions and decisions on the modeling of yield curves. For economists, the yield curve provides information about the risk taking of investors and the expectations of market participants.
There are three key theoretical bond market constructs that are related to yield curve analysis, which are the discount curve, the forward rate curve and the yield curve.
$P(\tau)$ denotes the current price of a bond if one is to receive 1 Euro in τ periods, i.e. the maturity of the bond is τ. $y(\tau)$ denotes the continuously compounded yield to maturity. The relationship between price and yield is defined by:

$$P(\tau) = \exp(-\tau\, y(\tau))\,, \tag{2.1}$$

representing the discount curve. Discount curve and yield curve $y(\tau)$ are directly linked and provide complete information about the other. From the discount curve one can derive the forward rate curve which is defined as:

$$f(\tau) = \frac{-P'(\tau)}{P(\tau)}\,. \tag{2.2}$$

The presented equations can be condensed to show the relationship between the yield curve and the forward rate curve:

$$y(\tau) = \frac{1}{\tau}\int_0^\tau f(u)du\,. \tag{2.3}$$

Theoretically, these three bond market constructs are interconvertible, because having information about one of them enables one to derive both the other curves. This thesis focuses solely on the yield curve, since most of the relevant literature for forecasting works with yield curves. In practice, yields cannot be directly observed, but the prices of traded bonds featuring different time periods to maturity are observable. (Diebold and Rudebusch, 2013). Yield construction, the practice of estimating yields from actually observed bond prices, is not discussed in this thesis.
There are several basic facts known about yield curves.

1. The average yield curve increases with maturity and is concave,

2. volatilities of yields decrease with maturities,

3. yields have a strong persistency which can be shown by autocorrelation, exemplary from one to 12 months,

4. yield spreads, the cross-sectional difference between yields of different maturities, show less volatility and persistency than the actual yields.

Yield curves can evolve to all kinds of shapes like upward sloping, downward sloping, humped and inverted humped. (Diebold and Rudebusch, 2013, p. 5-7; Diebold and Li, 2006, p. 343). Chapter 4 Description of data used further revisits these facts. This thesis analyzes zero-coupon government bond yield curves of Germany and the United States of America.

3 Methods and Models

This thesis presents and compares the performance of three different methods for modeling and forecasting yield curves. The first model presented was introduced by Diebold and Li (2006). They developed a new approach as they did not use the hitherto popular no-arbitrage models or the equilibrium models. The DL model is a three factor model for modeling and forecasting yield curves and has been widely used since its publication. It is therefore often used as a benchmark for the performance of other forecasting models (see for example Bowsher and Meeks (2008), Koopman et al. (2010), Klüppelberg and Sen (2010), Hays et al. (2012), Chen and Niu (2014)). The second method uses FPCA, a highly used technique for Functional Data Analysis (FDA). While FPCA is a general method for the analysis and modeling of functional data, this thesis refers to its application for functional time series forecasting following Hyndman and Shang (2009). The last model presented in this thesis applies Gaussian Processes and a dynamic modeling approach proposed by Sambasivan and Das (2017) for forecasting yield curves.

3.1 Nelson-Siegel framework and factor model by Diebold and Li

Diebold and Li built on the Nelson-Siegel yield curve as developed by Nelson and Siegel in 1987 and extended in 1988. They adjusted the commonly used framework with the focus on forecasting performance.

3.1.1 Exponential components framework by Diebold and Li

The Nelson-Siegel forward rate curve is described by:

$$f_t(\tau) = \beta_{1t} + \beta_{2t}\exp(-\lambda_t\tau) + \beta_{3t}\lambda_t\exp(-\lambda_t\tau). \tag{3.1}$$

It approximates a forward rate curve by a constant plus a polynomial multiplied by an exponential decay term[1]. The ensuing yield curve of the Nelson-Siegel model is as follows:

$$y_t(\tau) = \beta_{1t} + \beta_{2t}\left(\frac{1-\exp(-\lambda_t\tau)}{\lambda_t\tau}\right) + \beta_{3t}\left(\frac{1-\exp(-\lambda_t\tau)}{\lambda_t\tau} - \exp(-\lambda_t\tau)\right). \tag{3.2}$$

Exponential decay of the curve is regulated by the parameter λ_t, while small values generate slow decay, large values lead to fast decay. Additionally, the parameter controls where the loading on β_{3t} reaches its maximum.
DL introduced here a different factorization than the original formulation by Nelson and Siegel,

[1] Laguerre function.

which was as follows:

$$y_t(\tau) = b_{1t} + b_{2t}\left(\frac{1-\exp(-\lambda_t\tau)}{\lambda_t\tau}\right) - b_{3t}\exp(-\lambda_t\tau). \tag{3.3}$$

Both factorizations are equivalent and transferable by setting $b_{1t} = \beta_{1t}$, $b_{2t} = \beta_{2t} + \beta_{3t}$ and $b_{3t} = \beta_{3t}$. However, DL recognized the advantage gained by the new factorization which was better interpretability of the individual factors of the model. Due to the similar shape of $(1 - \exp(-\lambda_t\tau)/\lambda_t\tau)$ and $\exp(-\lambda_t\tau)$ in 3.3 the interpretation of the different factors and their estimation with the ensuing multicolinearity would render problematic.
One merit of the work of DL is consequentially the interpretation of the factors β_{1t}, β_{2t} and β_{3t}. Regularly, the three factors had been called long-term, medium-term and short-term factors respectively. DL, interpreting them as latent dynamic factors, built on this notion; The loading on β_{1t} is constantly 1, representing the long-term factor. The factor β_{2t} has the loading $\left(\frac{1-\exp(-\lambda_t\tau)}{\lambda_t\tau}\right)$. This function starts at 1 and decreases fast and monotonically to 0 - it is called the short-term factor. At last, the loading on β_{3t}, $\left(\frac{1-\exp(-\lambda_t\tau)}{\lambda_t\tau} - \exp(-\lambda_t\tau)\right)$, starts out at 0, reaches its maximum and decreases again to 0. Starting at 0, it is not short-term, and decaying to 0, it is not long-term, so it represents the medium-term factor. Figure 3.1 shows the different loadings dependent on maturity τ.

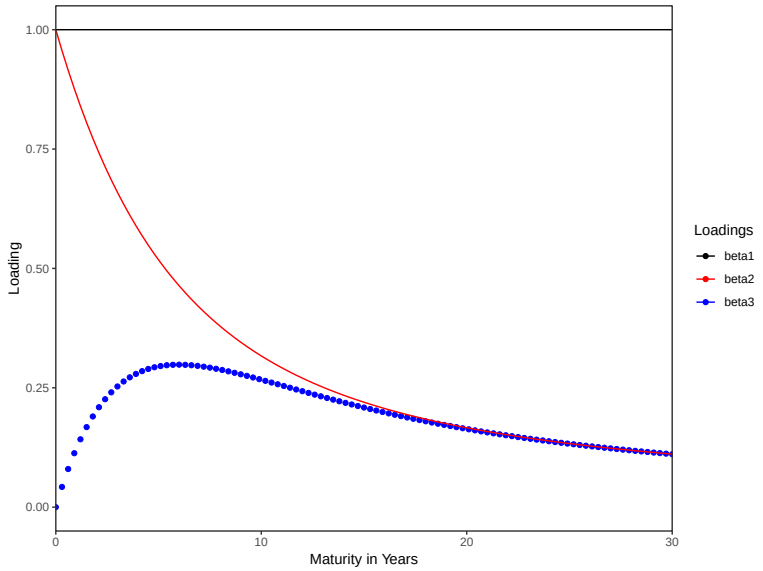


Figure 3.1: Factor loadings, exemplary with fixed $\lambda_t = 0.34$

DL offer another interpretation of the three factors relating them to level, slope and curvature, respectively. The first factor β_{1t} determines the level of the yield curve. Varying the loading would change all the yields in the same way, the loading of β_{1t} being the same for all maturities. Thus the change in β_{1t} would change the level of the entire yield curve. Also, $\lim_{\tau\to\infty} y_t(\tau) = \beta_{1t}$. The factor β_{2t} is connected to the slope of the yield curve. DL define the slope as the yield of the highest maturity in their data subtracted by the yield of the lowest maturity yield. Based on their data and with maturities of yields $y_t(\tau)$ stated in months, exemplary they show that

$y_t(120) - y_t(3) = -0.78\beta_{2t} + 0.06\beta_{3t}$. This indicates the great impact of β_{2t} on the slope. Another perspective is that short rates load more powerfully on β_{2t}. Varying this factor has a stronger influence on the yields at short maturities compared to long maturities, by this means making an impact on the yield curve slope.
At last, DL connect the factor β_{3t} to the curvature of the yield curve. They define the curvature as twice the yield of 2 years subtracted by the 10-year and 3-month yields. Based on their data it follows, that $2y_t(24) - y_t(120) - y_t(3) = 0.00053\beta_{2t} + 0.37\beta_{3t}$, showing the impact of β_{3t} on the curvature. (Diebold and Li, 2006, p. 340-343).
DL also demonstrate that the proposed model is able to reproduce different shapes of yield curves being upward or downward sloping or featuring a hump complying with the stylized facts. (Diebold and Li, 2006, p. 347).

3.1.2 Fitting the model and forecasting

In order to fit the model 3.2 the parameters need to be estimated at every step in t. For estimating $\{\beta_{1t}, \beta_{2t}, \beta_{3t}, \lambda_t\}$, nonlinear least squares method would be required, since 3.2 is not linear in λ_t. However, DL set λ_t at a fixed value $\lambda_t = \lambda$, thereby enabling the estimation to be done through ordinary least squares (OLS). In particular, the factor loadings, respectively regressors, of β_{2t} and β_{3t} can be computed and the β-vector is estimated by OLS. This approach improves computation by substituting a large number of optimization calculations with linear regressions and leads overall to a more simple and convenient handling. Concerning the choice of a fixed λ DL rely on the characteristic of λ_t to control the maturity where the loading on factor β_{3t} reaches its maximum. According to DL, usually two- and three-year maturities are selected to this end, so they chose the mean yielding the 30 month maturity. To maximize the loading on β_{3t} at a 30-month maturity λ_t must be $\lambda_t = 0.0609$, as stated by DL.[2]

Then the regression model is:

$$y_t(\tau_j) = \beta_{1t} + \beta_{2t}\left(\frac{1-\exp(-\lambda_t\tau_j)}{\lambda_t\tau_j}\right) + \beta_{3t}\left(\frac{1-\exp(-\lambda_t\tau_j)}{\lambda_t\tau_j} - \exp(-\lambda_t\tau_j)\right) + \epsilon_{jt}. \quad j = 1, \ldots, m. \tag{3.4}$$

The disturbances $\epsilon_{1t}, \ldots, \epsilon_{mt}$ are assumed tobe independent with mean zero and constant variance σ^2 for time t. (Koopman et al., 2010)
The estimates of the β-vector are understood as time series by DL and can hence be forecasted. The DL model uses autoregressive models on the factors of the model. Applying the AR(1) model and setting $\lambda_t = \lambda$, the forecasting equation is as follows:

$$\hat{y}_{t+h}(\tau) = \hat{\beta}_{1,t+h} + \hat{\beta}_{2,t+h}\left(\frac{1-\exp(-\lambda\tau)}{\lambda\tau}\right) + \hat{\beta}_{3t+h}\left(\frac{1-\exp(-\lambda\tau)}{\lambda\tau} - \exp(-\lambda\tau)\right), \tag{3.5}$$

with

$$\hat{\beta}_{i,t+h} = \hat{c}_i + \hat{\gamma}_i\hat{\beta}_{it}, \quad i = 1, 2, 3, \tag{3.6}$$

[2] Molenaars et al. (2015) correctly point out that this is not correct; Own calculations yield that λ_t must be 0.0598 to maximize at 30 months.

where $\hat{c}_i$ and $\hat{\gamma}_i$ result from regressing $\hat{\beta}_{it}$ on an intercept and on $\hat{\beta}_{i,t-h}$.
In this modeling setup the estimated yield curve only depends on $\{\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3\}$, so forecasting the curves is performed by forecasting the parameters $\{\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3\}$ recursively.

3.1.3 Excursus: AR-processes

Considering the auto regressive (AR) model used by DL, as background univariate time series models are described in the following. A time series can be described as "a sequence of observations in chronological order" (Ruppert and Matteson, 2015, p. 307) and comprises observations of a certain variable made at different points in time. Modeling times series enables forecasting of variables that are subject to temporal interrelations.
A time series process describes the notion of a time series variable realising different values. Therefore, it can be described as the sample space of a time series and an observed time series is sampled from this process. The examined variables are assumed to be random, hence, the process is referred to as a stochastic process. An important concept when analysing times series is the property of stationarity. Stationarity describes the behaviour of the time series across time and supposes that it behaves stochastically independently of the selected point in time. This form of strict stationarity is a very restrictive assumption, but the concept of weak stationarity is more commonly used. The following conditions for a stochastic process y_t apply:

$$E(y_t) = \mu, \qquad Var(y_t) = \sigma^2, \qquad \text{for all } t, \tag{3.7}$$

$$Cov(y_t, y_s) = \gamma(|t-s|), \qquad \text{for all } t \text{ and } s \text{ and a function } \gamma(h). \tag{3.8}$$

If these conditions are fulfilled, the first and second moments of the stochastic process y_t are independent of the point in time at which they are observed, hence are time invariant. Mean and variance are then constant and the autocovariance of two observations depends on their lag, that is the time span between them. The autocovariance function (ACVF) is denoted by γ, h being the lag: $\gamma(h) = \gamma(t+h, t)$. The autocorrelation function (ACF) ρ is defined as $\rho(h) = \gamma(h)/\gamma(0)$.

A univariate autoregressive process models the variable Y_t as a weighted sum of observations adding a disturbance term. An AR(1) is represented by:

$$y_t = \phi\, y_{t-1} + \epsilon_t, \tag{3.9}$$

ϵ_t is an i.i.d. random variable with mean μ and variance σ^2 and ϕ is a constant. ϵ_t is a white noise (WN) process, which is a stationary process often assumed to be normally distributed with mean 0 and variance σ^2. In order to analyse deviations from a central value the AR(1) process can be centralised by subtracting the mean of the process, which yields:

$$y_t - \mu = \phi\,(y_{t-1} - \mu) + \epsilon_t, \tag{3.10}$$

where $y_t - \mu$ has mean zero.
The term y_{t-1} in 3.9, respectively $y_{t-1} - \mu$ in 3.10, can be described as the memory of the process,

where a past value of the process feeds into the present value. This interrelation specifies the AR process an a correlated process. The parameter ϕ controls the strength of the relation between the past and the present value.
In extension to the AR(1) process, an AR process with p lags is presented. In an AR(p) process, p past values feed into the present value of the process. Considering the centralised AR(1) process in 3.10, the AR(p) process is defined by:

$$y_t - \mu = \phi_1 (y_{t-1} - \mu) + \phi_2 (y_{t-2} - \mu) + \ ... \ + \phi_p (y_{t-p} - \mu) + \epsilon_t, \tag{3.11}$$

where $\epsilon_t \sim \text{WN}(0, \sigma^2)$. The linear structure of the model replicates a linear regression model. Replacing the term $(\mu - \phi_1 \mu - \phi_2 \mu - \ ... \ - \phi_p \mu)$ with β_0 leads to the following:

$$y_t = \beta_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ ... \ + \phi_p y_{t-p} + \epsilon_t. \tag{3.12}$$

β_0, as $\beta_0 = (\{1 - \phi_1 - \phi_2 - \ ... \ - \phi_p \}\mu)$, can be interpreted as an intercept. Therefore, equation 3.12 is a multiple regression model with the past values of the time series as independent variables and can be estimated as such. (Ruppert and Matteson, 2015, chap. 12).
In order to select the optimal model to apply on an available data set an information criterion is required. One example, which is used frequently, is the Akaike information criterion (AIC). The AIC allows for selecting the order p of an autoregressiv model.(Ruppert and Matteson, 2015, chap. 5.12).

3.2 Functional Data Analysis

The concept for the second model that is presented in this thesis is a technique of FDA. FDA provides the framework to access a data set with the notion of functional interrelation between observations instead of perceiving observations only as individual data pairs. Characteristic of functional data is first the replication, measuring the same entity repeatedly, and a certain smoothness. The observed curves are thought of as items in themselves beyond the sequentially recorded observations. Although data is recorded at discrete points (y_j, x_j), y_j is understood to represent an instance of the assumed function at point x_j, capturing the intrinsic relation of the y_ms. (Ramsay and Silverman, 2005, pp. 1, 38).
In the context of yield curve modeling successive yield curves can then be understood as *functional time series*, where the yields recorded at different maturities are a discrete sampling of a true yield curve function. (Hays et al., 2012). In order to work with FDA methods the data has to fulfil certain characteristics and assumptions. As mentioned above, requisite for applying FDA approaches is the assumption of observed discrete data as being observations of a continuous process and that they generally are not subject to errors in measurement. (Hall et al., 2006). If one assumes the data to be recorded without error, no further smoothing is required, but interpolation in order to obtain a function from the observed values. (Ramsay and Silverman, 2005).

3.2.1 Functional principal component analysis

One of the main analytical tools of FDA is FPCA. With FPCA, functional processes can be characterized by their mean function and the eigenfunctions of the autocovariance operator. Hyndman and Shang (2009) use FPCA to forecast functional time series and propose extensions to this framework like weighted functional principal component regression. In the following the basic theory of FPCA and the application of weights as proposed by Hyndman and Shang is presented.

Originally, for FDA it is assumed that the examined functions are independent and identically distributed. This does not hold generally for financial data as for yield curve data examined in this thesis, but it is assumed there exists a relation between the sampled curves across time. An example of independent functions to apply FPCA on are temperature charts of hospitalized patients, where every observed curve or function represents an individual patient. Clearly, these functions are indendent when assuming that the patients' course of disease is independent. However, subsequent yield curves are not independent across time. This implied process structure has to be accounted for when applying FPCA. Klüppelberg and Sen (2010) and Hyndman and Shang (2009) approached this issue by applying autoregressive processes within the FPCA framework.

There are functions $f_1,\dots f_n$, sampled from a process f where f is assumed to be element of the Hilbert space $\mathcal{H} := L^2(T)$, provided the inner product $\langle f,g\rangle_{\mathcal{H}} = \int_T f(\tau)g(\tau)d\tau$ and the norm $\|f\| = \sqrt{\langle f,f\rangle_{\mathcal{H}}}$. It is assumed that observations are made on sufficient grid points τ_{ij} on domain T. Observations are measured with additive error W_{ij} presumed to be independent of the function generating process. For $i = 1,\dots,n$ and $j = 1,\dots,m$ measurements are:

$$\tilde{f}_i(\tau_{ij}) = f_i(\tau_{ij}) + W_{ij}, \quad E(W_{ij}) = 0,\, Var(W_{ij}) = \sigma^2. \tag{3.13}$$

The mean function of $f(\tau)$ and the continuous covariance function are defined by:

$$\mu_f = E(f(\tau)), \qquad \phi_f(\upsilon,\tau) = Cov(f(\upsilon), f(\tau)), \qquad \upsilon,\tau \in T. \tag{3.14}$$

Key for the application of FPCA is to interpret ϕ_f as the kernel of a linear mapping on the Hilbert space $L^2(T)$, as $\phi_f : \mathcal{H} \mapsto \mathbb{R}$. With α this operator is defined as follows:

$$(\phi_f\alpha)(\upsilon) = \int_T \phi_f(\upsilon,\tau)\,\alpha(\tau)\,d\tau, \qquad \alpha \in L^2(T), \tag{3.15}$$

yielding a function in υ, with notation for operator and kernel being the same. Covariance function is defined by:

$$\phi_f(\upsilon,\tau) = \mathbb{E}\left[(f(\upsilon) - \mu_f(\upsilon))\,(f(\tau) - \mu_f(\tau))\right]. \tag{3.16}$$

(Benko et al., 2009).

θ_k are the eigenvalues of the covariance operator ϕ_f, ordered by $\theta_1 \geq \theta_2 \geq \ldots \geq 0$. ψ_ks are the respective eigenfunctions of the operator. The eigenfunctions of the operator form a complete orthonormal sequence on $L^2(T)$. Applying the Karhunen-Loève decomposition (also expansion) yields a representation of individual curves of function f:

$$f(\tau) = \mu_f(\tau) + \sum_{k=1}^{\infty} \xi_k \psi_k(\tau), \tag{3.17}$$

with

$$\xi_k = \int_T (f(\tau) - \mu_f(\tau))\, \psi_k(\tau) d\tau \tag{3.18}$$

is the k-th so-called functional principal component score with $\mathbb{E}(\xi_k) = 0$ and $\mathrm{Var}(\xi_k) = \lambda_k$. Klüppelberg and Sen (2010), Hall et al. (2006).
Considering model equation 3.13, it yields that

$$\mathbb{E}(\tilde{f}_i(\tau_{ij})) = \mu_f(\tau), \quad Cov(\tilde{f}_i(\upsilon), \tilde{f}_i(\tau)) = \phi_f(\upsilon, \tau) + \sigma^2 I(\upsilon = \tau). \tag{3.19}$$

The mean function μ_f and the covariance operator ϕ are estimated from observed data. Curves f_i can be approximated by replacing estimates and applying only a certain number K of eigenfunctions in 3.17. The number of eigenfunctions can be chosen by the portion of variance explained. (Klüppelberg and Sen, 2010, p. 3f.).

3.2.2 Estimation of functional principal components

To obtain eigenfuctions and FPCA scores, first the mean function and covariance function need to be estimated. This is conducted by comprising the available sample of curves and smoothing the generated scatterplot. For each i, data pairs $\{(\tau_j, \tilde{f}_{ij}),\, i = 1, \ldots, n,\, j = 1, \ldots, m\}$ are observed. A local linear model is estimated for the non parametric regression of $\tilde{f}$ on τ. This implies finding $\hat{\beta}_0(s)$ and $\hat{\beta}_1(s)$, with $s \in T$,

$$\sum_{i=1}^{n} \sum_{j=1}^{m} \{\tilde{f}_{ij} - \beta_0(s) - \beta_1(s)\,(\tau_j - s)\}^2\, K_1\left(\frac{\tau_j - s}{b_f}\right). \tag{3.20}$$

b_f is the selected smoothing bandwidth (can be selected by generalized cross-validation). K_1 is a compactly supported symmetric univariate kernel function, that needs to be square integrable and endowed with a finite variance and absolutely integrable Fourier transform.
The estimated mean function is then set $\hat{\mu}_f(s) = \hat{\beta}_0(s)$.
For estimating the covariance functions of the functional process f one estimates the covariance surface ϕ_f. All pairwise empirical covariances of 3.16 $\phi_i(\tau_{j1}, \tau_{j2}) = (\tilde{f}_{ij1} - \hat{\mu}_f(\tau_{j1}))\,(\tilde{f}_{ij2} - \hat{\mu}_f(\tau_{j2}))$ are comprised into a scatterplot with $\{[(\tau_{j1}, \tau_{j2}),\, \phi_i(\tau_{j1}, \tau_{j2})],\, i = 1, \ldots, n,\, j_1, j_2 = 1, \ldots, m\}$. For smoothing, a nonparametric regression of $\phi_i(\tau_{j1}, \tau_{j2})$ on (τ_{j1}, τ_{j2}) is fitted finding

the minimizers $\hat{\beta}_0(s_1, s_2)$, $\hat{\beta}_1(s_1, s_2)$, $\hat{\beta}_2(s_1, s_2)$:

$$\sum_{i=1}^{n} \sum_{1 \leq j_i \neq j_2 \leq m} \{\phi_i(\tau_{j1}, \tau_{j2}) - [\beta_0(s_1, s_2) + \beta_1(s_1, s_2)(s_1 - \tau_{j1}) + \beta_2(s_1, s_2)(s_2 - \tau_{j2})]\}^2 \times K_2 \left(\frac{\tau_{j1} - s}{h_f}, \frac{\tau_{j2} - s}{h_f} \right). \tag{3.21}$$

smoothing bandwidth h_f can be selected like b_f. K_2 now is a square integrable and compactly supported radially symmetric bivariate kernel function endowed with a finite variance and absolutely integrable Fourier transform.
The estimated covariance surface is $\hat{\phi}_f(s_1, s_2) = \hat{\beta}_0(s_1, s_2)$. The estimates for eigenvalues λ_k and eigenfunctions ψ_k, $\{\hat{\lambda}_k, \hat{\psi}_k\}$, are obtained by solving the eigenequations

$$\int \hat{\phi}_f(\upsilon, \tau)\hat{\psi}_k(\upsilon)d\upsilon = \hat{\lambda}_k\hat{\psi}_k(\tau). \tag{3.22}$$

Orthonormality constraints on $\hat{\psi}_k$ and constraints to positive definiteness on the covariance surface apply. To estimate the first K scores of FPCA 3.18 by a discrete integral approximation

$$\hat{\xi}_k = \sum_{j=2}^{m} (\tilde{f}_{ij} - \hat{\mu}_f(s_{ij}))\,(s_{ij} - s_{i,j-1})\hat{\psi}_k(sij), \quad i = 1, \ldots, n,\ k = 1, \ldots, K. \tag{3.23}$$

Individual functions are represented by the empirical version of 3.17 for selected number of eigenfunctions K:

$$\hat{f}_i(\tau) = \hat{\mu}_f(\tau) + \sum_{k=1}^{K} \hat{\xi}_{ik}\hat{\psi}_k(\tau). \tag{3.24}$$

(Klüppelberg and Sen (2010), Müller et al. (2006)).

3.2.3 Functional principal component model for forecasting

Hyndman and Shang present the following model for their implementation of the FPCA:

$$y_i(\tau_j) = f_i(\tau_j) + \sigma_i(\tau_j)\epsilon_{ij}, \tag{3.25}$$

where $i = 1, \ldots, n$ and $j = 1, \ldots, m$. ϵ_{ij} are white noise with unit variance and $\sigma_i(\tau)$ is time-dependent and can thereby reflect heteroskedasticity. They apply nonparametric smoothing separately on every observed curve to generate estimates of $f_i(\tau)$ to be estimated by a FPCA model:

$$f_i(\tau) = \mu(\tau) + \sum_{k=1}^{K} \xi_{ik}\psi_k(\tau) + e_i(\tau), \tag{3.26}$$

with eigenfunction $\psi_k(\tau)$ as the kth principal component function and $\{\xi_{1k}, \ldots, \xi_{nk}\}$ are the kth principal component scores. $e_i(\tau)$ are independent and identically distributed (i.i.d.) random functions with zero mean.
Eigenfunctions ψ_k and ψ_l are orthogonal for $k \neq l$ which implies that the principal component scores ξ_{ik} are uncorrelated. They can then be forecasted applying univariate time series models.

The respective scores ξ_{ik}, $k = 1, \ldots, K$ are interpreted as univariate time series and can then be forecasted separately. The forecasted principal component scores that are obtained this way are then multiplied with the principal component functions producing predictions of curves. Further, Hyndman and Shang propose geometrically decreasing weights in the principal component decomposition in order to increase the influence of the more recent data on the forecasts. While they did not apply their method on financial but on demographic data, this approach is evenly relevant in the context of financial forecasting, because on markets more recent developments have a greater influence than occurences longer ago.
The weights are accounted for in the mean function $\mu(\tau)$ by computing a weighted average of the estimated smoothed functions $f_i(\tau)$:

$$\hat{\mu}_\tau = \sum_{i=1}^{n} w_i \, \hat{f}_i(\tau), \tag{3.27}$$

with $\hat{f}_i(\tau)$ being the smoothed curve estimated from $\tilde{f}_i(\tau)$. The weights $w_i = \kappa(1-\kappa)^{n-i}$ with $0 < \kappa < 1$ are geometrically decreasing from the most recent data curve to the data curves in the past. Naturally, if weights are not supposed to be used, w_i is a vector of ones. When using weights, κ can be derived empirically by minimizing the mean integrated forecast error (MISFE):

$$\text{MISFE}(h) = \int_{\tau_1}^{\tau_m} (f_{n+h}(\tau) - \hat{f}_{n+h|n}(\tau))^2 \, dx. \tag{3.28}$$

To forecast functions, drawing on 3.25 and 3.26 and replacing estimates, it follows:

$$f_i(\tau_j) = \hat{\mu}(\tau_j) + \sum_{k=1}^{K} \xi_{ik}\psi_k(\tau_j) + \hat{e}_i(\tau_j) + \hat{\sigma}_i(\tau_j)\hat{\epsilon}_{ij}. \tag{3.29}$$

The principal component scores ξ_{ik} are forecasted by a univariate time series model and conditioning on observed data $\boldsymbol{\mathcal{I}} = \{y_t(\tau_j) \, : \, i = 1, \ldots, n, \; j = 1, \ldots, m\}$ and principal components $\boldsymbol{\Psi} = \{\psi_1(\tau), \ldots, \psi_k(\tau)\}$ h-step ahead forecasts are represented by:

$$\hat{y}_{i+h|i}(\tau) = \mathbb{E}[y_{i+h}(\tau) | \, \boldsymbol{\mathcal{I}}, \boldsymbol{\Psi}] = \hat{\mu}(\tau) + \sum_{k=1}^{K} \hat{\xi}_{i+h|i,k} \, \psi_k(\tau) \tag{3.30}$$

(Hyndman and Shang, 2009).

3.3 Gaussian Processes

The last approach to model yield curves that is presented is based on GP. While GP have been known for a long period of time, possibly appearing as early as late 19th century, in the 1990s they became known by a broader audience in the machine learning context. They can be applied to supervised learning problems (Rasmussen and Williams, 2006). The Gaussian process model is introduced in the following.
A Gaussian process is a random process that is defined by its mean and covariance matrix $\boldsymbol{\Sigma}$. $f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n)$ are understood as random variables, given the input data $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}\}$. These

variables follow a joint multivariate normal, respectively Gaussian, distribution. For simplicity reasons in this introduction a process with zero-mean is assumed:

$$f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_n) \sim \boldsymbol{N}(\boldsymbol{0}, \boldsymbol{\Sigma}) \tag{3.31}$$

Exemplary, the covariance $\boldsymbol{\Sigma}_{pq}$ between $f(\boldsymbol{x}_p)$ and $f(\boldsymbol{x}_q)$ is a function of $\boldsymbol{x}_p$ and $\boldsymbol{x}_q$. This covariance function $\mathbf{K}$ is chosen to generate a positive definite covariance matrix. For $f(\boldsymbol{x}_p)$ and $f(\boldsymbol{x}_q)$ the covariance is defined as $\boldsymbol{\Sigma}_{pq} = \mathbf{K}(\boldsymbol{x}_p, \boldsymbol{x}_q)$. The choice of a covariance function is a modeling decision. A widely used covariance function for the multi-dimensional case is the squared exponential function:

$$Cov\left[f(\boldsymbol{x}_p), f(\boldsymbol{x}_q)\right] = \mathbf{K}(\boldsymbol{x}_p, \boldsymbol{x}_q) = \nu \, \exp\left[-\frac{1}{2}\sum_{d=1}^{D} w_d \left(\boldsymbol{x}_p^d, \boldsymbol{x}_q^d\right)\right]^2 + \nu_0, \tag{3.32}$$

with $\boldsymbol{x}_p^d$ denoting the dth component of an D-dimensional vector $\boldsymbol{x}_p$. Parameter ν controls the vertical scale of variation and the w_ds control the horizontal lengthscale. (Ažman and Kocijan, 2005).

Gaussian processes can be embedded in the Bayesian modeling framework. Contrary to the frequentist approach to parameter learning like for example the OLS estimation for the β-vector in the DL model, which assumes constant parameters to be estimated, Bayesian statistics assumes that an unknown parameter is a random variables with a probability distribution. Knowledge about this distribution is termed prior. Based on the likelihood of a sample of the data and the prior a posterior distribution can be derived which is used for estimation. (Sambasivan and Das, 2017).

Now, the following model is assumed:

$$y = f(\boldsymbol{x}) + \boldsymbol{\epsilon}, \tag{3.33}$$

$\boldsymbol{\epsilon}$ is white noise with $\boldsymbol{\epsilon} \sim \boldsymbol{N}(0, \nu_0)$. A GP prior with covariance function 3.32 and unknown parameters is induced on $f(.)$.

$y_1, \ldots, y_N, y_{N+1} \sim \boldsymbol{N}(\boldsymbol{0}, \mathbf{K}_{N+1})$, where $\mathbf{K}(\boldsymbol{x}_p, \boldsymbol{x}_q) = \boldsymbol{\Sigma}_{pq} + \nu_0 \delta_{pq}$. If $p = q$ then $\delta_{pq} = 1$, otherwise $\delta_{pq} = 0$. Now, $y_1, \ldots, y_N, y_{N+1}$ is divided into two parts, $\boldsymbol{y} = [y_1, \ldots, y_N]$ and $y_* = y_{N+1}$. It follows that

$$\boldsymbol{y}, y_* \sim \boldsymbol{N}(\boldsymbol{0}, \mathbf{K}_{N+1}) \tag{3.34}$$

and

$$\mathbf{K} = \begin{bmatrix} [\mathbf{K}] & [\mathbf{k}(\mathbf{x}_*)] \\ [\mathbf{k}(\mathbf{x}_*)^T] & [\boldsymbol{k}(\mathbf{x}_*)] \end{bmatrix}, \tag{3.35}$$

with $\mathbf{K}$ as an $N \times N$ matrix that gives covariances between y_p and y_q of x_p and x_q respectively for $p, q = 1, \ldots, N$. $\mathbf{k}(\mathbf{x}_*)$ gives the covariances between y_* and y_p, i.e. $\mathbf{k}(\mathbf{x}_*) = \mathbf{K}(\mathbf{x}_*, \mathbf{x}_p)$ in a $N \times 1$ for $p = 1, \ldots, N$. Lastly, $\boldsymbol{k}(\mathbf{x}_*) = \mathbf{K}(\mathbf{x}_*, \mathbf{x}_*)$ is the covariance between the function of the new input and itself.

From this joint probability of $\boldsymbol{y}, y_*$ a marginal and a conditional distribution can be derived. Provided a training sample of N data pairs $\{\mathbf{x}_p, \mathrm{y}_p\}$, $p = 1, \ldots, N$, the marginal distribution yields the likelihood of the observations $\mathbf{y}|\mathbf{X} \sim \boldsymbol{N}(\mathbf{0}, \mathbf{K})$, with $\mathbf{y}$ being training target values and $\mathbf{X}$ the associated training in put values.
By maximizing the likelihood, respectively technically easier the log-likelihood, of $\mathbf{y}|\mathbf{X}$ the unknown parameters of the covariance function and the parameter for the noise variance ν_0 can be estimated. The conditional distribution yields the prior distribution of y_* given the new input values $\mathbf{x}_*$, when conditioning the joint distribution on the training sample and the new input values $\mathbf{x}_*$: $p(y_*|\mathbf{y}, \mathbf{X}, \mathbf{x}_*) = p(\mathbf{y}, y_*)/p(\mathbf{y}|\mathbf{X})$. This distribution is Gaussian with the following mean and variance:

$$\mu(\mathbf{x}_*) = \mathbf{k}(\mathbf{x}_*)^T \mathbf{K}^{-1} \mathbf{y} \tag{3.36}$$

$$\sigma^2(\mathbf{x}_*) = \boldsymbol{k}(\mathbf{x}_*) - \mathbf{k}(\mathbf{x}_*)^T \mathbf{K}^{-1} \mathbf{k}(\mathbf{x}_*) + \nu_0. \tag{3.37}$$

Ažman and Kocijan (2005), Williams (1997).

3.3.1 Gaussian process regression model

Building on the ideas presented in the preceding section, the approach by Sambasivan and Das (2017) is presented in the following. They show in their paper that GP can be applied to the forecasting of yield curves using Gaussian Process Regression. To that end, Sambasivan and Das formulate the following model:

$$f(\tau) = \mu(\tau) + W(\tau), \tag{3.38}$$

$$\boldsymbol{y} = \mu(\tau) + W(\tau) + \boldsymbol{\epsilon}. \tag{3.39}$$

The function $\mu(\tau)$ is a parametric function and the process $W(\tau) \sim \boldsymbol{N}(\mathbf{0}, \mathbf{K})$. $\mathbf{K}$ is the Covariance function of f.
With m as the number of observed data points ,

$$f \sim \boldsymbol{N}_m(\mu(\tau), \mathbf{K}), \tag{3.40}$$

$$\epsilon \sim \boldsymbol{N}_m(0, \sigma_\epsilon^2 \boldsymbol{I}_m), \tag{3.41}$$

$$\boldsymbol{y} \sim \boldsymbol{N}_m(f(\tau), \mathbf{K} + \sigma_\epsilon^2). \tag{3.42}$$

Derived from the likelihood function $L(f|\boldsymbol{y}, \boldsymbol{\phi}, \sigma^2)$ the negative log-likelihood function $l(f)$ and the corresponding negative log-posterior function $p(f)$ are as follows:

$$l(f) \propto \frac{1}{2\sigma_\epsilon^2} \, (\boldsymbol{y} - f)^T [\mathbf{K} + \sigma_\epsilon^2 \boldsymbol{I}]^{-1} (\boldsymbol{y} - f), \tag{3.43}$$

$$p(f) \propto \frac{1}{2\sigma_\epsilon^2} \left((\boldsymbol{y} - f)^T [\mathbf{K} + \sigma_\epsilon^2 \boldsymbol{I}]^{-1} (\boldsymbol{y} - f) + f^T \mathbf{K}^{-1} f \right). \tag{3.44}$$

In order to make estimations the posteriori of the function is computed via Bayes theorem. To estimate y for a new given input-vector τ_*, one samples the functions from the posterior and

computes the mean value at τ_*.

$$\hat{f}(\tau_*) = E(f|\tau_*, \boldsymbol{y}) \tag{3.45}$$
$$= \mu(\tau_*) + \mathbf{K}(\tau_*, \tau)[\mathbf{K}(\tau, \tau) + \sigma_\epsilon^2.I]^{-1}(\boldsymbol{y} - \mu(\tau)). \tag{3.46}$$

Also, the variance of the estimate at τ_* can be given by:

$$Var(f_*) = \mathbf{K}(\tau_*, \tau_*) - \mathbf{K}(\tau_*, \tau)[\mathbf{K}(\tau, \tau) + \sigma_\epsilon^2 \boldsymbol{I}]^{-1}\mathbf{K}(\tau, \tau_*). \tag{3.47}$$

The following section elaborates on further on the forecasting method.

3.3.2 Dynamic Gaussian Process prior model for forecasting

Sambasivan and Das formulate a dynamic GP prior model on which they base their forecasting procedure:

$$\boldsymbol{y}_i = \mu_i(\tau) + \boldsymbol{\epsilon}_i, \tag{3.48}$$

with $\boldsymbol{y}_i = \begin{pmatrix} y_i(\tau_1) \\ y_i(\tau_2) \\ \vdots \\ y_i(\tau_m) \end{pmatrix}$ and $\boldsymbol{\epsilon}_i = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \end{pmatrix}$, containing the terms to all maturities m.

$\mu_i(\tau)$ depicts the mean function and its system equation is defined as follows:

$$\mu_i(\tau) = \mu_{i-1}(\tau) + W_i. \tag{3.49}$$

W_i is a process described by $W_i(\tau) \sim \boldsymbol{N}_m(\mathbf{0}, \mathbf{K}_{i-1})$ with:

$$\mathbf{K}_{i-1} = \mathbf{K}(\tau, \tau'|\rho_{i-1}), \tag{3.50}$$

where ρ_{i-1} is the parameter estimated at time point $i-1$ controlling $\mathbf{K}$.
Now, when data $\boldsymbol{Y}_i = (\boldsymbol{y}_i, \boldsymbol{y}_{i-1}, \boldsymbol{y}_{i-2}, \ldots, \boldsymbol{y}_1)$ is known, inference about μ_i and forecasts of $\boldsymbol{y}_{i+1}$ is possible by applying Bayes theorem. This idea is critical to the understanding of the proposed procedure. Bayes theorem in this context can be stated as follows:

$$\mathbb{P}(\mu_i(\tau)|\, \boldsymbol{Y}_i) \;\propto\; \mathbb{P}(\boldsymbol{y}_i|\, \mu_i, \boldsymbol{Y}_{i-1}) \times \mathbb{P}(\mu_i(\tau)|\, \boldsymbol{Y}_{i-1}), \tag{3.51}$$

showing the posterior process of $\mu(\tau)$ as being proportional to the likelihood times the prior process of $\mu(\tau)$.
At time $i-1$ the posterior process of $\mu(\tau)$ is distributed with posterior mean function $\hat{\mu}_{i-1}(\tau)$ and the posterior covariance function $\hat{\mathbf{K}}_{i-1}$ of the process at time $i-1$:

$$\mu_{i-1}|\, \boldsymbol{Y}_{i-1} \;\sim\; \boldsymbol{N}_m\Big(\hat{\mu}_{i-1}(\tau), \hat{\mathbf{K}}_{i-1}\Big)\,. \tag{3.52}$$

Then at time point i the prior predictive process is:

$$\mu_i|\, \boldsymbol{Y}_{i-1} \;\sim\; \boldsymbol{N}_m\Big(\hat{\mu}_{i-1}(\tau), \hat{\mathbf{K}}_{i-1}\Big)\,, \tag{3.53}$$

the likelihood function and marginal likelihood function respectively are as follows:

$$\boldsymbol{y}_i|\, \mu_i(\tau), \boldsymbol{Y}_{i-1} \;\sim \boldsymbol{N}_m\big(\mu_i(\tau), \sigma^2_{i-1}\boldsymbol{I}_m\big)\,, \tag{3.54}$$

$$\boldsymbol{y}_i|\, \boldsymbol{Y}_{i-1} \;\sim \boldsymbol{N}_m\Big(\hat{\mu}_{i-1}(\tau), \hat{\mathbf{K}}_{i-1} + \sigma^2_{i-1}\boldsymbol{I}_m\Big)\,. \tag{3.55}$$

Sambasivan and Das introduce hyperparameters $\{\rho_{i-1}, \sigma_{i-1}\}$ of the covariance function, which are estimated by maximizing the marginal likelihood 3.55. Parameter ρ controls the covariance function $\mathbf{K}$, σ^2 is an error variance associated with σ^2_ϵ in 3.41.
Following the original paper, the term hyperparameter refers to a parameter known from a prior distribution in terms of Bayesian optimization contrary to the use of the term in a machine learning context where hyperparameters have to be set before model training. Rasmussen and Williams (2006) indicate, that they are parameters of a nonparametric model, therefore termed hyperparameters.
The observation at time point i can be forecasted by using the expected value of $\boldsymbol{y}_i|\, \boldsymbol{Y}_{i-1}$ from 3.55:

$$\begin{aligned}\hat{\mu}_i(\tau_*) &= \mathbb{E}\left(\boldsymbol{y}_i(\tau_*)|\, \boldsymbol{Y}_{i-1}\right)\\ &= \mathbf{K}(\tau_*, \tau|\hat{\rho}_{i-1})[\mathbf{K}(\tau, \tau|\hat{\rho}_{i-1}) + \hat{\sigma}^2_{i-1}\boldsymbol{I}]^{-1}\boldsymbol{y}_{i-1}(\tau).\end{aligned} \tag{3.56}$$

Posterior covariance function and posterior mean function are then updated at every forecasting step after observing $\boldsymbol{y}_i$, yielding respectively:

$$\hat{\mathbf{K}}_{i.updated} = \mathbf{K}(\tau_*, \tau_*|\hat{\rho}_i) - \mathbf{K}(\tau_*, \tau|\hat{\rho}_i)\left[\mathbf{K}(\tau, \tau|\hat{\rho}_i) + \hat{\sigma}^2_{i-1}\boldsymbol{I}\right]^{-1}\mathbf{K}(\tau, \tau_*|\hat{\rho}_i) \tag{3.57}$$

and

$$\begin{aligned}\hat{\mu}_{i.updated} &= \mathbb{E}\left(\hat{\boldsymbol{\mu}}_i(\tau_*)|\, \boldsymbol{Y}_i\right)\\ &= \hat{\mu}_i + \mathbf{K}(\tau_*, \tau|\hat{\rho}_i)\left[\mathbf{K}(\tau, \tau|\hat{\rho}_i) + \hat{\sigma}^2_i\boldsymbol{I}\right]^{-1}\left(\boldsymbol{y}_i(\tau) - \hat{\mu}_i\right),\end{aligned} \tag{3.58}$$

of the updated posterior process over $\boldsymbol{y}_i$:

$$\boldsymbol{y}_i(\tau)|\, \boldsymbol{Y}_i \;\sim\; \boldsymbol{N}_m\Big(\hat{\mu}_{i.updated}, \hat{\mathbf{K}}_{i.updated}\Big)\,. \tag{3.59}$$

3.4 Discussion of proposed methods

The parametric model by DL was developed with the sole purpose of estimating and forecasting yield curves. Since its publication it has grown to be widely accepted and has been used as a benchmark model for further research. It comprises a statistical three-factor model assuming the yield curve to be a continuous function of maturity τ. Modeling parameter are $\{\beta_{1t}, \beta_{2t}, \beta_{3t}, \lambda_t\}$.

DL propose to set $\lambda_t = \lambda$ to enable OLS estimation of the β-vector. Then, yield functions can be modeled with estimates of only the β-vector. DL understand the estimated coefficients, the $\boldsymbol{\beta}_j$-vector for $j = 1, 2, 3$, as a univariate time series which can be forecasted by autoregressive time series models.

Modeling functional time series using FPCA is a common approach. With this approach, yield curves are understood as functional time series. The observed yields at different maturities are a discrete sampling from a true underlying function generating yield curves. Individual curves can be estimated by an approximation based on the decomposition of the covariance function. Modeling parameters are hence derived from the structure of the estimated covariance functions of the observed functions. Hyndman and Shang additionally suppose that the more recently observed curves have a greater impact on the forecasts going forward than curves that have been observed earlier.

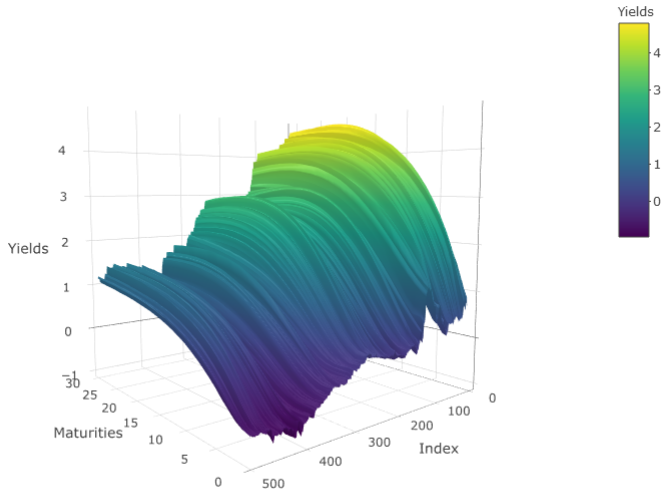
The nonparametric approach based on GP assumes a joint multivariate normal distribution on the target values, while the covariance between the target values depends on the inputs of the examined functions. The covariance determines the estimated mean function and covariance function used for defining the GP, from which the estimated curves are assumed to be sampled. The approach by Sambasivan and Das to use a dynamic gaussian process prior model on yield curves is a new proposition and has not been reviewed in the literature yet.

4 Description of data used

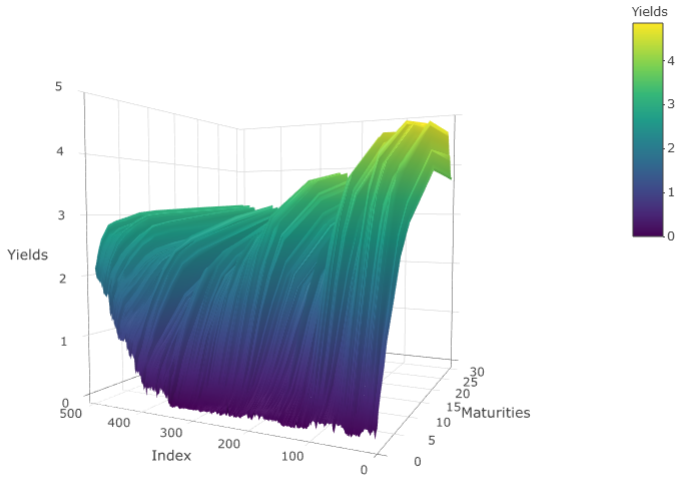
In order to test proposed methods two different data sets were used in this thesis containing yield curves for zero-coupon government bonds. One data set comprises data derived from market-listed *Bundeswertpapiere* of the Federal Republic of Germany. The data is published by the German Central Bank *Bundesbank* and can be downloaded from *https://www.bundesbank.de*. The other set comprises data derived from market quotations of Treasury securities by the United States. It is published by the US Department of the Treasury and data can be downloaded from *https://www.treasury.gov*.
If one wanted to observe the values of the yield curve on the market, the listing of a risk-free zero-coupon bond at every maturity would be required. Since there are very few such listings yield curve data has to be estimated from available bond data. One of the most prevalent methods for yield curve construction is the approach developed by Nelson and Siegel and extended by Svensson, which is also used by the German *Bundesbank*. (Deutsche Bundesbank, 1997) Both data sets were prepared in the same way hence the following description applies to the German and United States (of America) (US) data alike. The time span selected is 2009-04 till 2018-09 in order to obtain data post financial crisis. From the available daily data weekly observations from the beginning of each week were extracted. Observations with *NAs* for holidays and non-trading days were removed. Also, in the case of US yields a term of maturity that was not available for the entire selected time span was removed.
Additional smoothing is not required because avalaible data is not noisy data but data pairs of smoothed functions. Characteristic of yield curve data as functional data is, that the domain does not change within the original data set. Values τ_{ij} are the same for every observed curve i and functional observations y_{ij} are provided. The data is non-periodic (Ramsay and Silverman, 2005).

Both data sets comprise 496 periods and hence 496 yield curves. For *Bundesbank* data 31 terms of maturities are available: {6 months, 1 year, 2 years, 3 years, ..., 30 years}. For US Treasury data 11 terms of maturities are available: {1 month, 3 months, 6 months, 1 year, 2 years, 3 years, 5 years, 7 years, 10 years, 20 years, 30 years}. An overview of the data sets are given in figure 4.1.

The US yield curves display changing shapes, particularly a flattening of the curves in 2017, with rising yields in the short term maturity regions (see also figure 4.1b). All the yield curves and the mean curve are displayed in figures 4.2 and 4.3. Mean curves of only part of the available data illustrate this change of shapes in comparison with figure 4.4. For the US yield curves considerably less terms of maturities are available and they are more unevenly spaced than



(a) Weekly yield curves of German *Bundesbank* from 2009 till 2018 (Index) at 31 maturities.



(b) Weekly yield curves of US Treasury from 2009 till 2018 (Index) at 11 maturities.

Figure 4.1: Overview of used data sets.

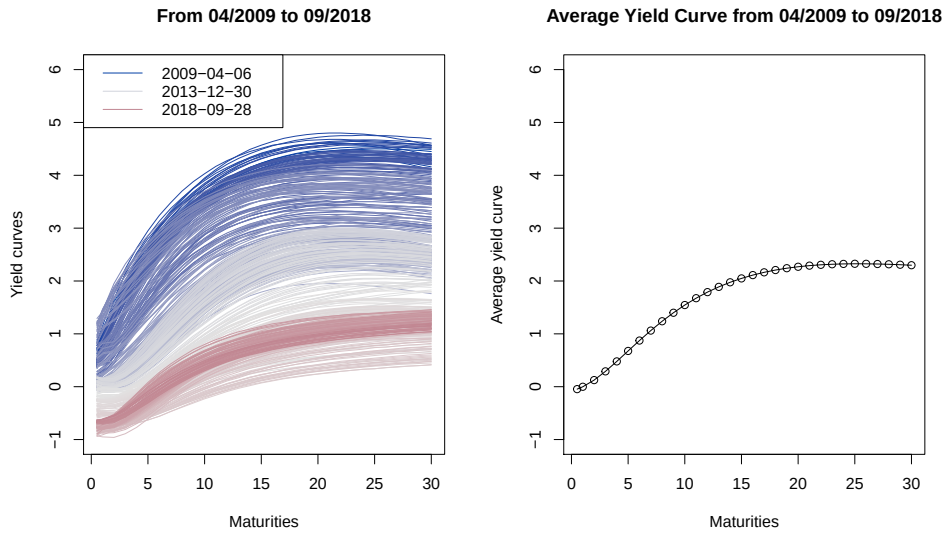


Figure 4.2: Overview of yield curves and mean by German Central Bank.

Bundesbank (BB) data. This supposedly impacts the forecasting performance negatively.
It shows that the yield curve comply with the basic facts about yield curves as for example the mean curves increase with maturity and are concave.

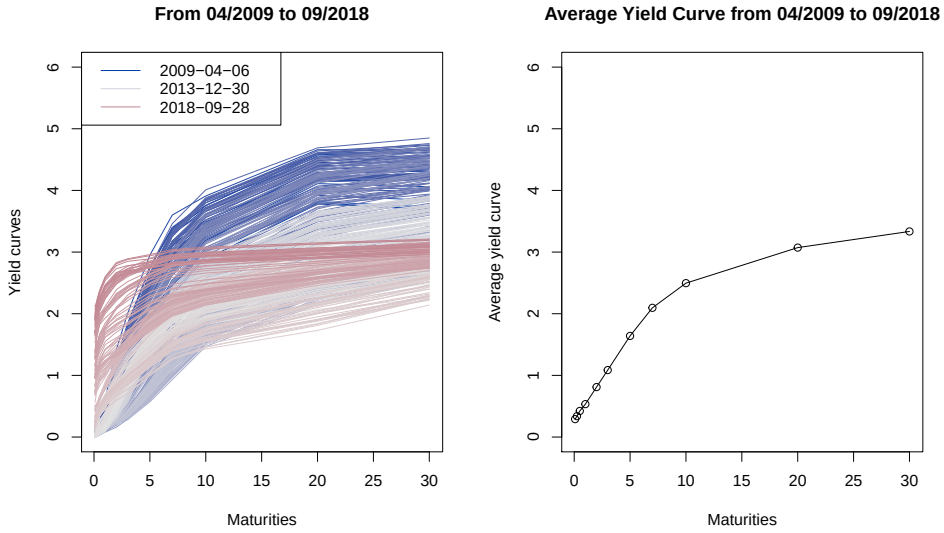


Figure 4.3: Overview of yield curves and mean by US Treasury.

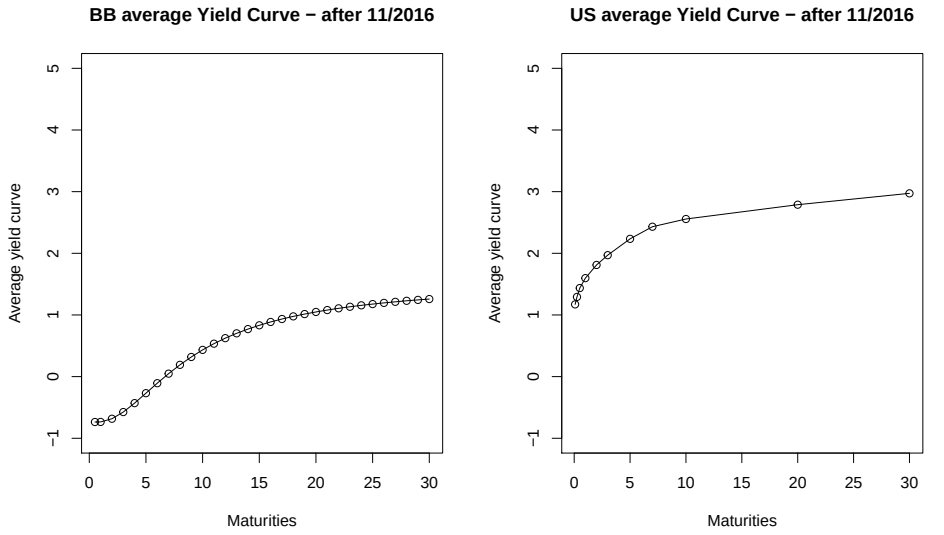


Figure 4.4: Mean curves of part of the data sets.

5 Modeling and forecasting

This chapter is dedicated to the empirical evaluation of the different modeling methods for yield curves discussed in chapter 3. To this end a general study design was set up that was applied to every modeling method described in the subsequent section. Forecasting performance was evaluated and compared by a particular measurement of prediction accuracy further detailed in section 5.2.
First, every modeling method is recapitulated and the forecasting approach to be applied to the data is described. In the results sections the used variables are discussed and the results of the respective modeling set-ups are presented. At the end of this chapter forecasting performance of all methods are compared.
All empirical analyses are conducted using R version 3.5.1 (R Core Team, 2018) and Python 3.6. The used code and data sets can be found in the electronic appendix.

5.1 Description of study design

The focus of this thesis was to analyze the forecasting performance of different models. In order to obtain robust results that are based on single experiments a particular study design was applied. This approach is two-fold based on two methods, the out-of-sample testing and cross-validation. Out-of-sample testing means that the available data is partitioned into a training and a testing or holdout sample. Model fitting is only carried out on the training sample and forecasting and evaluating forecasting performance is executed on the testing sample. This way, the forecasting procedure is tested on data that was not used for fitting the model. (Bergmeir et al., 2018, also basis for illustration). The following figure illustrates the procedure with blocks representing yield curve data, showing the training data with filled blocks (blue), testing data as striped (orange) and data that is not used in white. Exemplary forecasting horizon is 4 periods.

Figure 5.1: Illustration of out-of-sample forecasting.

The second method uses K-fold cross-validation to apply the forecasting procedure to K different training samples. By this means a more valid conclusion about the forecasting performance of a certain model can be drawn than basing conclusions only on a single modeling set-up. The observed data, however, cannot be randomly split in training and testing samples, due to the temporal relation in the data. (Bergmeir et al., 2018, also basis for illustration). Therefore, a rolling window of the training data with specified window length l was used producing multiple forecasting models. Based on the current window of data predictions for the subsequent h

periods are made and the mean forecasting error over the entire curves is recorded. Stepwise moving through the entire data set repeating this prediction process yields a measure for prediction accuracy for this method based on the set size respectively number of weeks of the chosen rolling window. The figure 5.2 illustrates this approach with the stepwise shifting of the window.

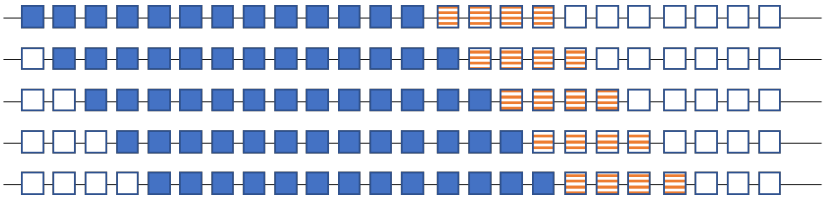


Figure 5.2: Illustration of forecasting procedure with rolling window.

For every presented method a standard set-up of the model is defined. For the DL and the FPCA model also variations of this standard set-up are evaluated. For DL this concerns the lag of the AR(p) process used to forecast the modeling factors, for the FPCA model weights are introduced as an extension to the standard model. The GP is evaluated by the standard model.

Another aspect this thesis is examining is the question how many observed periods are required to gain valuable information for forecasting yield curves. To this end models are evaluated with different numbers of training periods to examine the influence on the forecasting performance. How this examination relates to the described forecasting procedure, is illustrated in figure 5.3.

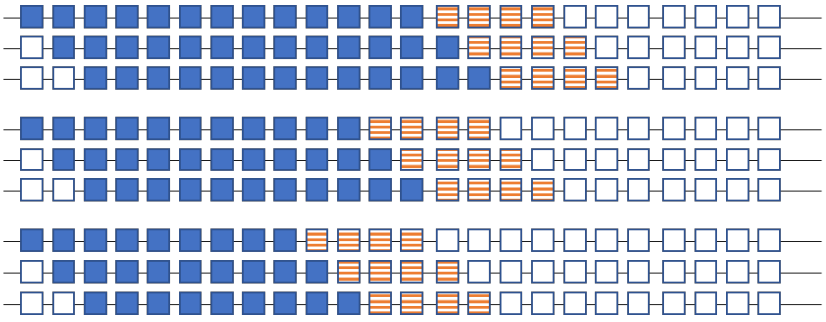


Figure 5.3: Illustration of varying size of training sample within out-of-sample and cross-validation testing framework.

As figure 5.3 exemplifies with three partitions, the forecasting study goes through the data as described in 5.2 shifting a training sample window of the same size. To examine the influence of the training window the size of this window is varied, decreasing by steps of two in this example. The step size by which the rolling window is shortened is selected in consideration of the forecast horizon h. The smaller the rolling window is, the more models can be estimated shifting through the data. Out-of-sample testing with smaller window size leads to a mean forecast error to be calculated over more estimated models. Overall, the forecasted periods respectively weeks the models were tested with are $h = 4$ und $h = 26$ to reflect a 1-month and 6-month forecasting horizon.

5.2 Measurement of prediction accuracy

In order to evaluate the forecasting performance the root mean squared error (RMSE) was used as measurement of prediction accuracy. RMSE is the square root of the mean squared error and the RMSE of a forecasted curve f_i is defined by:

$$\text{RMSE}_i = \sqrt{\frac{\sum_{\tau=1}^{m}(\hat{y}[\tau,i] - y[\tau,i])^2}{m}}, \tag{5.1}$$

with $\hat{y}[\tau,i]$ being the forecasted yield for period i and maturity τ, $y[\tau,i]$ being the actual yield at the respective period i and maturity τ. The RMSE of a forecast is then the mean of RMSE all forecasted yield curves.
RMSE is considered a "standard measure in the financial literature for measuring and comparing the accuracy of interest rates prediction models" (Arbia and Di Marcantonio, 2015) as it is also used by Diebold and Li (2006), Hays et al. (2012), Chen and Niu (2014) and Sambasivan and Das (2017).
To calculate RMSE the forecast error measurement function of the `"ftsa"` R package by Hyndman and Shang (2018), Shang (2013) was used on the methods implemented in R for this thesis.
The disadvantage of the RMSE, however, is that it is sensitive to the number and the spacing of available support points, as it only captures differences between the values at those support points. In contrast, error measurements based on integration better reflect the shape of the estimated function and do not depend on the observed points. Since yield curves are relatively smoothly shaped, in this thesis the RMSE is applied.

5.3 Factor model by Diebold and Li

Diebold and Li use the Nelson-Siegel factor model in the following form (see 3.2):

$$y_t(\tau) = \beta_{1t} + \beta_{2t}\left(\frac{1-\exp(-\lambda_t\tau)}{\lambda_t\tau}\right) + \beta_{3t}\left(\frac{1-\exp(-\lambda_t\tau)}{\lambda_t\tau} - \exp(-\lambda_t\tau)\right).$$

Estimating the loadings on factors $\beta_{1t}, \beta_{2t}, \beta_{3t}, \lambda_t$ forms the basis of forecasting yield curves following DL. For this thesis this is conducted with the `"YieldCurve"` R package by Guirreri (2015). Contrary to the approach described in the paper by DL, in this package the parameter λ_t is not fixed but is estimated at every step in t. Particularly, at every forecasting step iteratingly for every maturity a λ is estimated maximising the loading on β_{3t}. Then by OLS the *beta*s for this λ are estimated. The parameter set with those *beta*s minimizing the residuals of the estimation model across all maturities is then selected.
Figures 5.4 and 5.5 depict the sample autocorrelation of the three factors as estimated by the DL model. They show how all series display significant autocorrelation in the first lags which supports the proposition to forecast the factors by an autoregressive process.

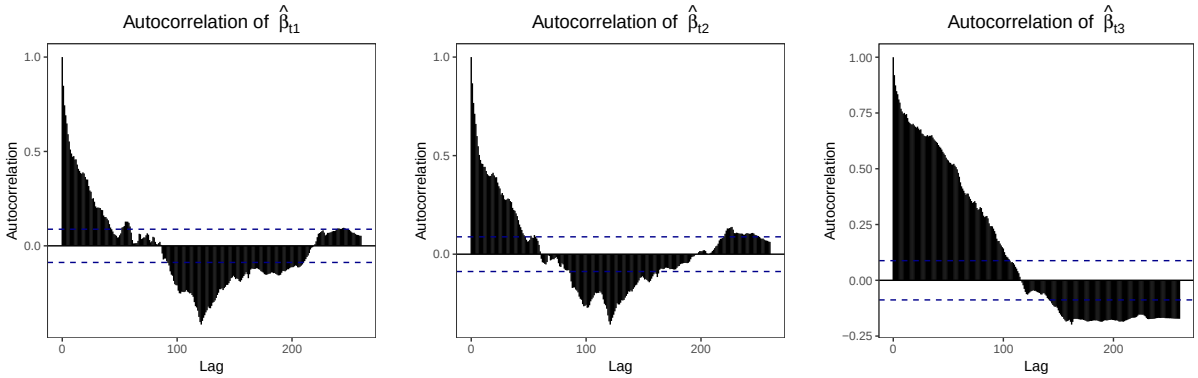


Figure 5.4: Sample autocorrelations of the estimated β-vector of *Bundesbank* data with lags in weeks, plotted with a 95%-confidence interval.

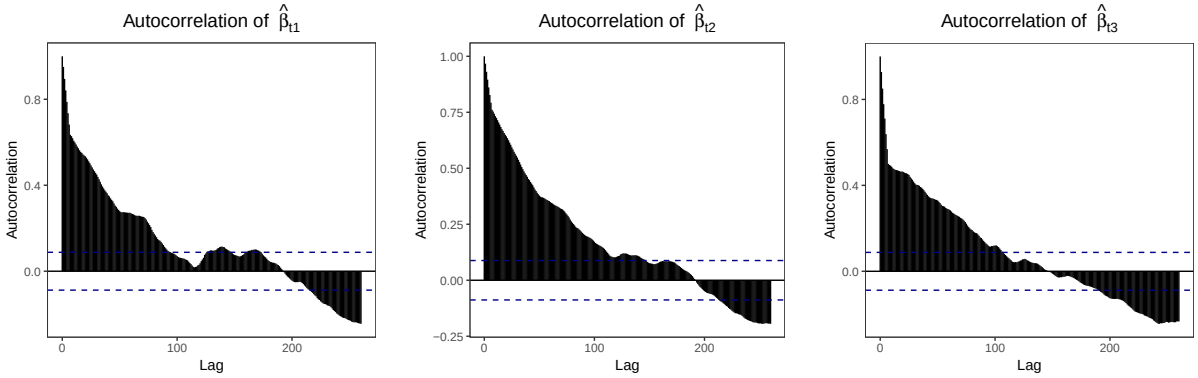


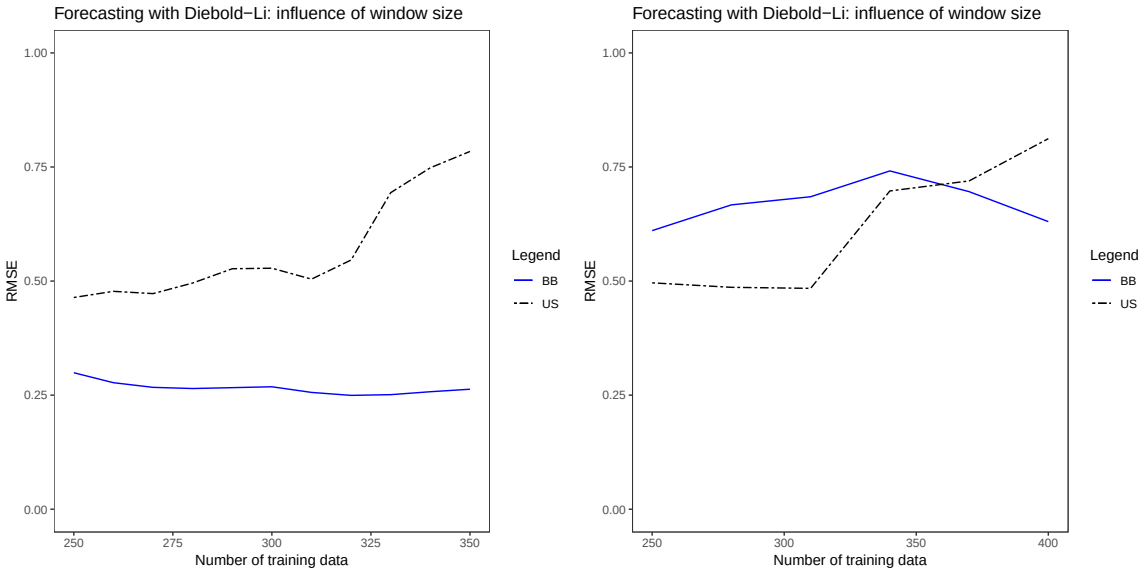
Figure 5.5: Sample autocorrelations of the estimated β-vector of US Treasury data with lags in weeks, plotted with a 95%-confidence interval.

5.3.1 Forecasting

For forecasting yield curves with the Nelson-Siegel the factors $\beta_{1t}, \beta_{2t}, \beta_{3t}$ are forecasted applying a univariate AR model. DL apply an AR(1) model, which they call "the simplest great workhorse [of] autoregressive models". In this thesis the lag of the AR process forecasting the univariate time series of the β-vector is identified as one variable to be further examined. Whether improvements in forecasting are achieved extending the AR model to a higher number of lags is discussed in the subsequent section of this chapter.
For comparison DL also apply a multivariate AR model to the factors by using a VAR(1) model. However, they caution against the use of VARs with regard to potential for overfitting and the questionable additional value due to little interaction across the factors and low correlation between them. Owing to a forecasting performance worse than that of the AR model the VAR is not considered in DL's concluding forecast accuracy comparison. Accordingly, in this thesis the VAR model is not considered. DL do not consider ARMA processes respectively moving average (MA) components for prediction.
To set a fixed λ_t for forecasting as postulated by DL the mean of all estimated λ_ts is calculated. This approach is also proposed by Arbia and Di Marcantonio (2015). Molenaars et al. (2015) found that the forecasting performance is relatively insensitive to the choice of λ_t. The forecasts are made recursively from 1-step to h-steps ahead from the end of the training sample of the observed time series. With this modeling setup predictions for different forecast horizons can be conducted.

5.3.2 Results

The standard model representative for the DL model to be examined in this thesis is forecasting the β-vector with an AR(1) process. The forecasting performance of the DL model was further evaluated in terms of the lag of the AR process used for forecasting the factors of the model. Forecasting was performed with forecast horizons of $h = 4$ and $h = 26$ weeks using both data sets presented in chapter 4.
Another question this thesis considers is how the choice of the size of the training data window influences the forecasting performance of the models. For this purpose the following figures in 5.6 display the variation of the RMSE dependent on length l of the rolling window of training data to generate a certain forecast. The figures show the mean RMSE resulting from forecasting successively models with the respective fixed window length l.
The window length of the training data has a relatively small impact on the RMSE with the modeling set-up of 250 to 350 periods in the training sample for $h = 4$ and 250 to 400 periods for $h = 26$. This holds true especially for the data set of the German BB. The US data shows more sensitivity to the number of training periods which reflects the changing shapes in the later part of the US data set, which is forecasted by a larger window containing regular and flatter shapes of curves.
To illustrate the forecasting performance of the different models boxplots depict the mean RMSE of forecasts with rolling windows. For the short forecast horizon of $h = 4$ RMSE displays a performance difference between the forecasts of BB and US data while the yield curves from BB



(a) Forecast horizon 4.

(b) Forecast horizon 26.

Figure 5.6: Influence of window size on RMSE using the DL model for multiple step ahead forecasts of *Bundesbank* (BB) and US Treasury (US) data.

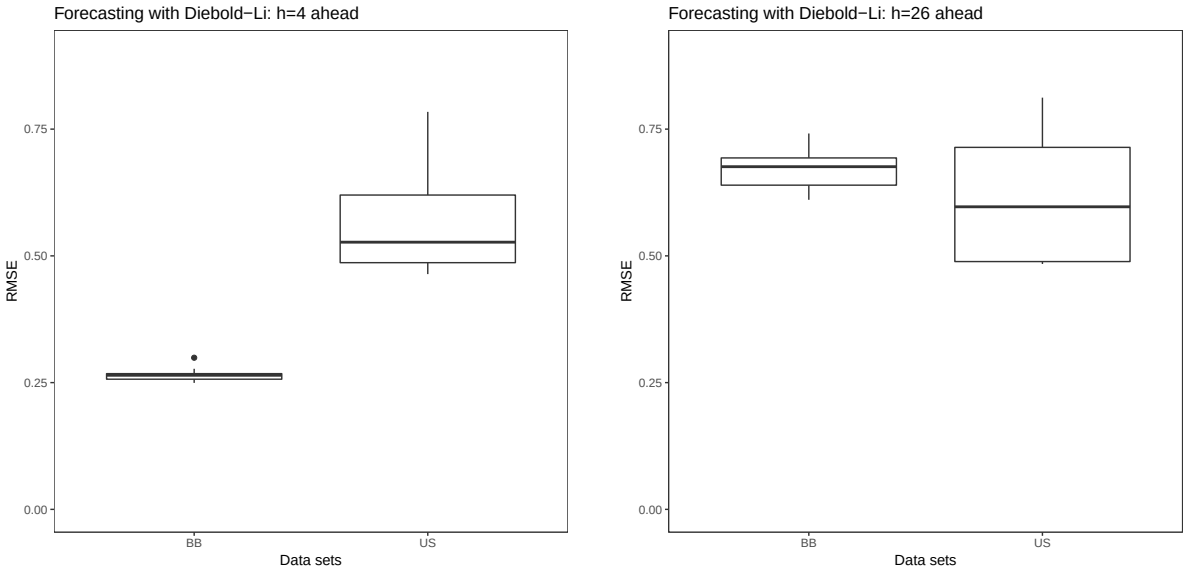


Figure 5.7: Boxplots of RMSE using the DL model for multiple step ahead forecast. Forecasts were conducted with 250 to 350 training periods for $h = 4$ and 250 to 400 for $h = 26$.

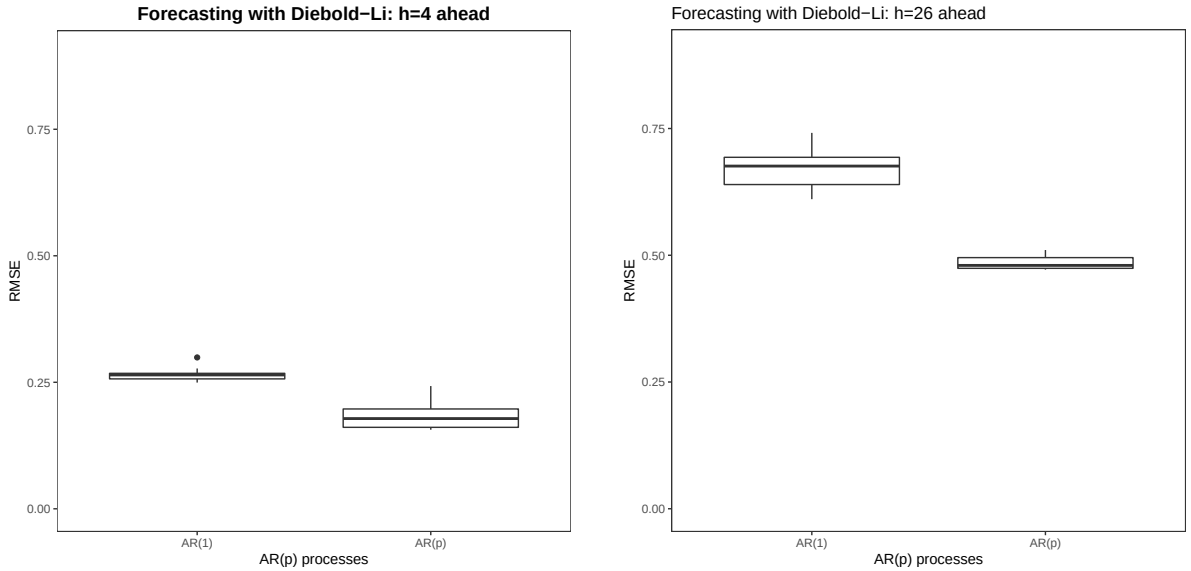


Figure 5.8: Boxplots of RMSE using the DL model for multiple step ahead forecasts of *Bundesbank* data. $\beta-$vector is forecasted with AR(1) respectively AR(p) processes.

are predicted better and with less variation. For the long forecast horizon $h = 26$ the difference between the data sets is smaller, while BB data still showing less variation.

As a variation of the original DL model this thesis also evaluates if applying an AR(p) process with a higher lag for forecasting yields better results. To this end the models presented in figure 5.7 are evaluated optimizing the AR(p) process up to a lag of 4, selecting the optimal model by AIC. While this extension improves the forecasts with BB data, the impact is not as clear with the US data.
The greater variance in forecast RMSE with US data compared to BB data is a consequence from the evolving shapes of the US data while the shape of the BB data changes relatively less across time.

5.4 Functional principal component model

We recall from equation 3.17 that each function $f_i - \mu$ can be represented with its generalized Fourier expansion in the eigenfunctions ϕ_ks, which yields

$$f_i(\tau) = \mu(\tau) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(\tau).$$

Hyndman and Shang proposed geometrically decreasing weights to emphasize the influence of the more recent data on the forecasts. The weights are accounted for in the mean function $\mu(\tau)$ by computing a weighted average as seen in equation 3.27:

$$\hat{\mu}_\tau = \sum_{i=1}^{n} w_i \, \hat{f}_i(\tau).$$

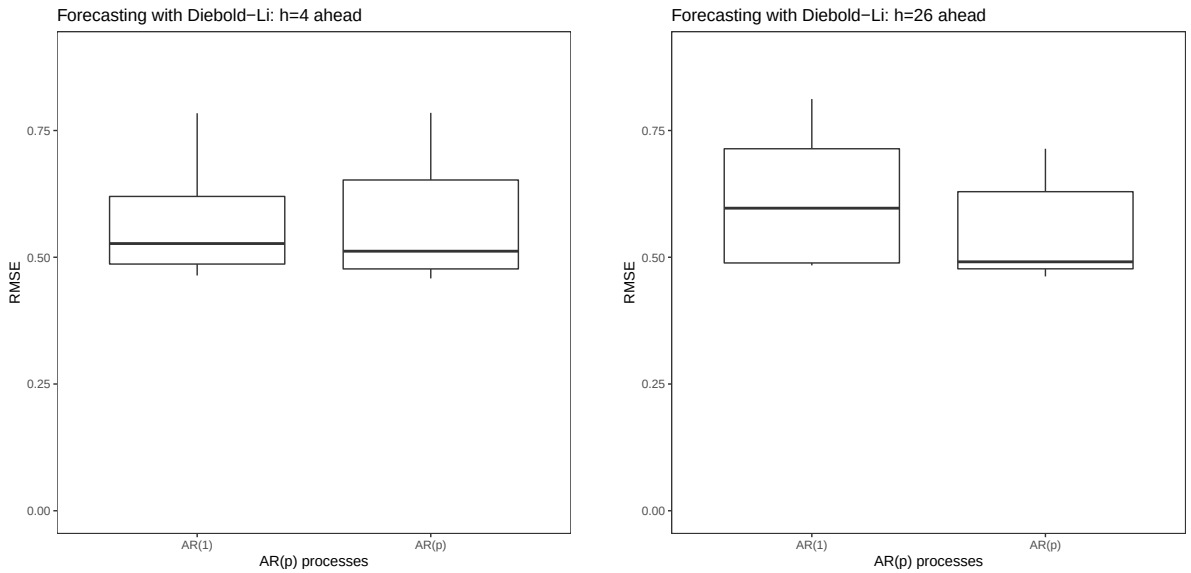


Figure 5.9: Boxplots of RMSE using the DL model for multiple step ahead forecasts of US Treasury data. $\beta-$vector is forecasted with AR(1) respectively AR(p) processes.

To obtain the principal component functions and scores, Hyndman and Shang apply the following approach: the smoothed functions $\hat{f}_i^*$ are discretized on a grid with q densely and equally spaced points $\{\tau_1^*, \ldots, \tau_q^*\}$ on the interval of $[\tau_1, \tau_m]$. This yields an $n \times q$ matrix $\boldsymbol{G}^*$ and it is defined that $\boldsymbol{G} = \boldsymbol{W}\boldsymbol{G}^*$ with $\boldsymbol{W} = \text{diag}(w_1, \ldots, w_n)$. Through singular value decomposition to $\boldsymbol{G}$ it follows $\boldsymbol{G} = \boldsymbol{\Phi}\boldsymbol{\Lambda}\boldsymbol{V}'$. Now, $\phi_k(\tau_r^*)$ is the (r,k)th element of $\boldsymbol{\Phi}$ with $r = 1, \ldots, q$. Assuming $\boldsymbol{B} = \boldsymbol{G}\boldsymbol{\Phi}$, ξ_{ik} is the (i,k)th element of $\boldsymbol{B}$. Further required values of $\phi_k(\tau)$ can be obtained via linear interpolation. (Hyndman and Shang, 2009).

5.4.1 Forecasting

For forecasting yield curves via FPCA the `"ftsa"` R package by Hyndman and Shang was used (Hyndman and Shang, 2018, Shang, 2013). It provides fitting of a principal component model to a functional time series object and forecasting of the FPCA scores applying univariate time series forecasting methods.
Influencing the principal component method is the selection of the number K of principal components to fit. For yield curves, however, only few eigenvalues are required making this assumption less essential. The results part of this chapter further elaborates on this matter.
With this modeling setup predictions for different forecasts horizons can be carried out.

5.4.2 Results

The standard model for the FPCA model does not apply weights and estimates the functions with a selected number K of eigenfunctions. In order to select the number of eigenfunctions required for forecasting yield curves different models were compared by mean RMSE. Exemplary, models with BB data, forecast horizon $h = 4$ and training sample sizes between 250 and 350 periods are evaluated. Across all examined models the smallest variation explained by the first principal components is not less than 97%. Further exemplary studies for FPCA models with

the short forecast horizon of $h = 4$ show that up to using $k = 3$ principal components the mean RMSE of the models still decreases. Since forecasts are not sensitive to the choice of the number of eigenfunctions K, if K is sufficiently large, however a small K might lead to poor forecasting performance, for the following analyses of the model K was set to 3. (See also Hyndman and Shang, 2009, p. 5). For US data on average the first principal component does not explain such a high portion of variance, but half of total variation, which is offset, however, by the second principal component, together explaining more than 90%.

Figure 5.10 and 5.11 show the (first) three functional principal components and the respective

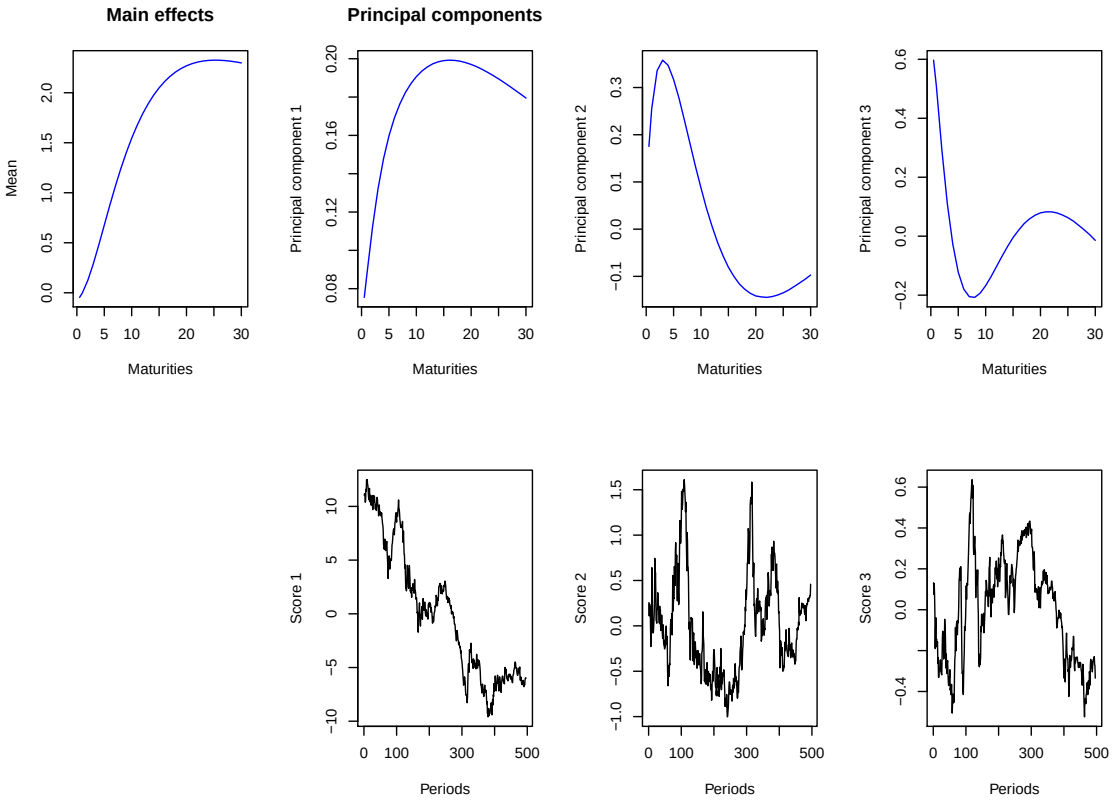


Figure 5.10: The first three weighted functional principal components and respective scores for *Bundesbank* data.

scores of the two data sets. Interpretation of functional principal components is straightforward, as the principal components represent "the major modes of variation" of the yield curves over the maturities (Benko et al., 2009). The shapes of the principal components comprises information about the shapes to be found in the data set, particularly if the number of principal components is small, as it is here (Hall et al., 2006). For the BB data the first principal component is similarly shaped to the mean but with a turning point at about 15 years (maturity) which illustrates the general shape of the yield curves becoming flatter at this point. A similar shape is displayed by the second principal component of the US data, while its first component resembles the second of BB data. The axis scaling of the scores reflect the decreasing impact on the estimation of $f_i(\tau)$ of the associated principal components.

In the reference paper for the FPCA model Hyndman and Shang introduce a weight vector in

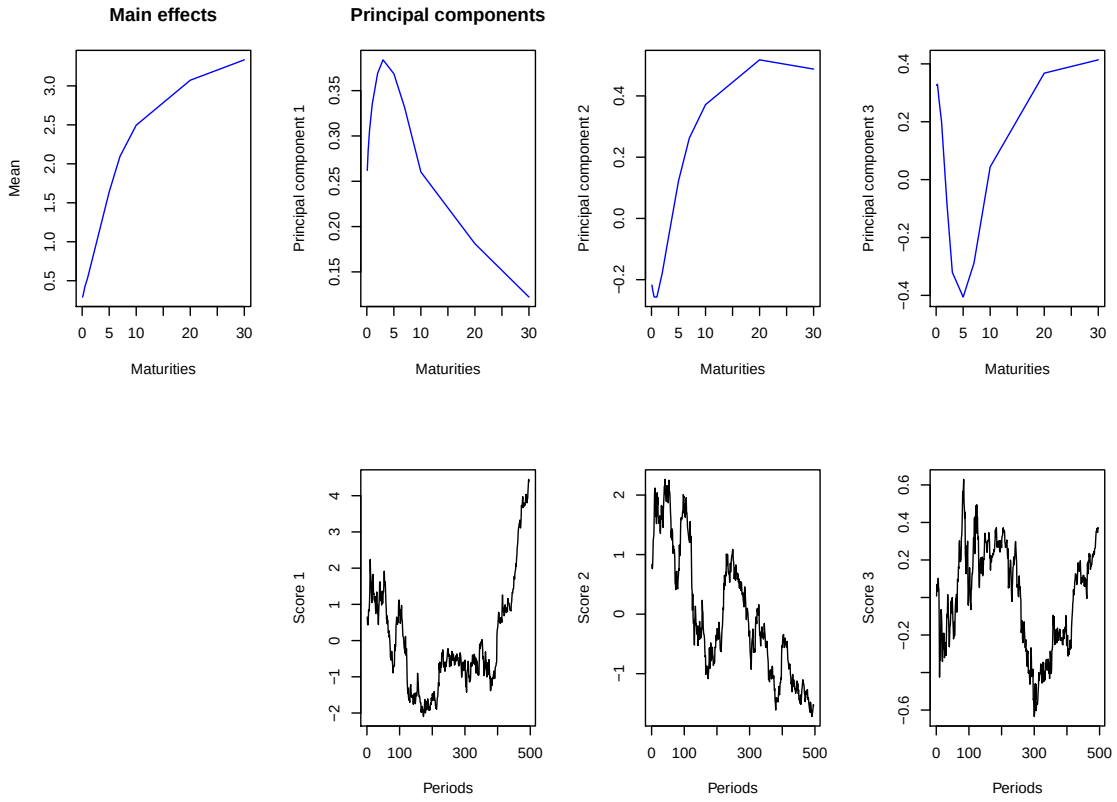
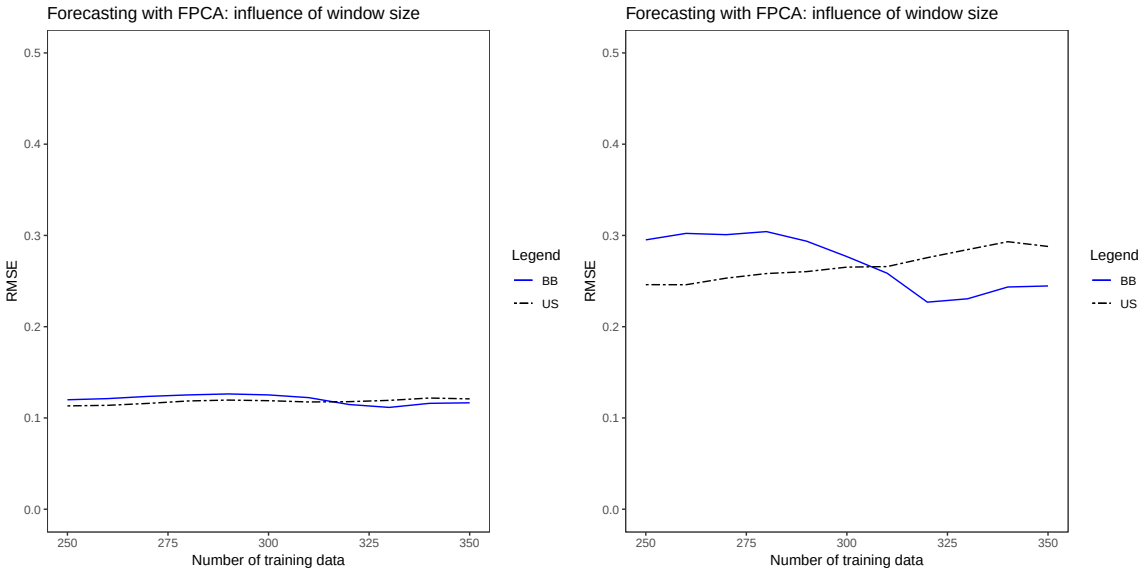


Figure 5.11: The first three weighted functional principal components and respective scores for US data.

the mean function showing this improves forecasting performance. Forecasting models without and with weights are compared in this section.
First, the influence of the length of the rolling window training sample is illustrated in figure 5.12. With FPCA modeling the mean RMSE is relatively insensitive to window length, particularly in the case of forecasting $h = 4$ steps.
The evaluation of the FPCA standard model as illustrated in figure 5.13 shows comparable results for BB and US data.

As an extension to the standard model weights are now applied to the mean function as proposed by Hyndman and Shang, which increase the influence of more recently observed curves on the current forecast. In figures 5.14 and 5.15 a comparison of mean RMSE between standard models without weight and the alternative models with weight is depicted. It shows that the weighting of the observations improves forecasting performance for US data while the result for BB data is not as clear. For the forecast horizon $h = 4$ only a very small improvement is generated, for the long horizon there is no improvement.



(a) Forecast horizon 4.

(b) Forecast horizon 26.

Figure 5.12: Influence of window size on RMSE using the FPCA model for multiple step ahead forecasts of *Bundesbank* (BB) and US Treasury (US) data.

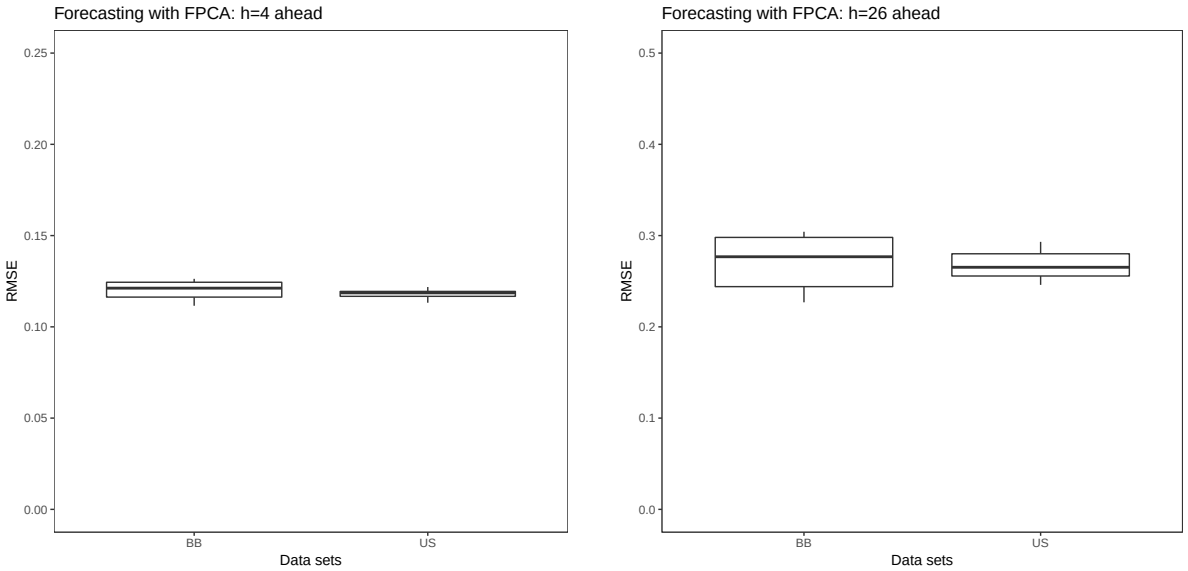


Figure 5.13: Boxplots of RMSE using the FPCA model for multiple step ahead forecast. Forecasts were conducted with 250 to 350 training periods for $h = 4$ and $h = 26$.

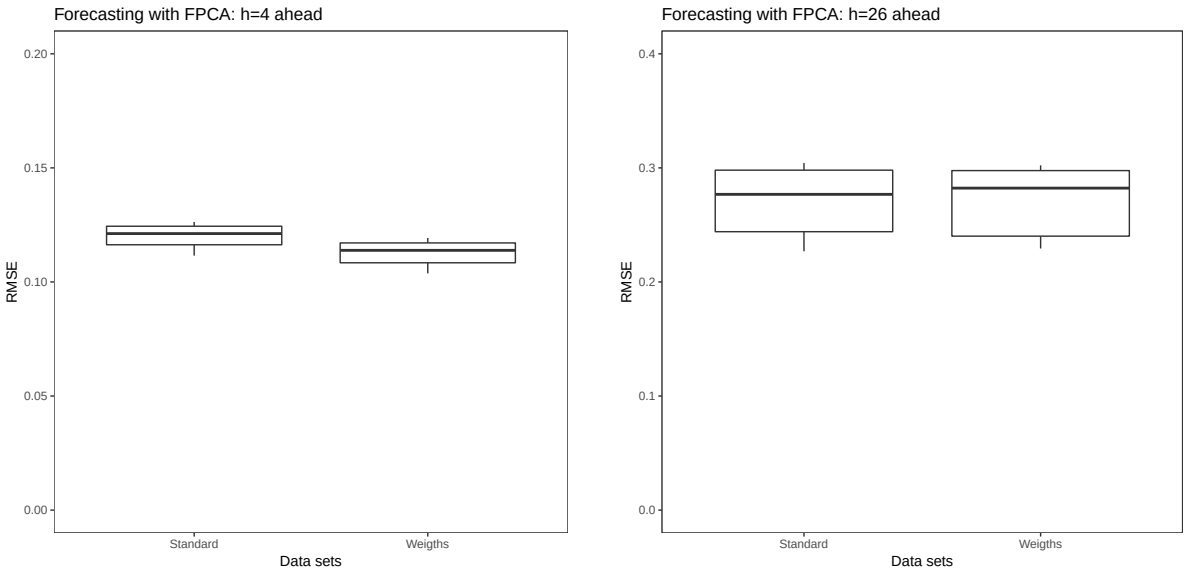


Figure 5.14: Boxplots of RMSE using the FPCA model for multiple step ahead forecasts of *Bundesbank* data. Forecasting is conducted without weights (Standard) respectively with weights (Weights).

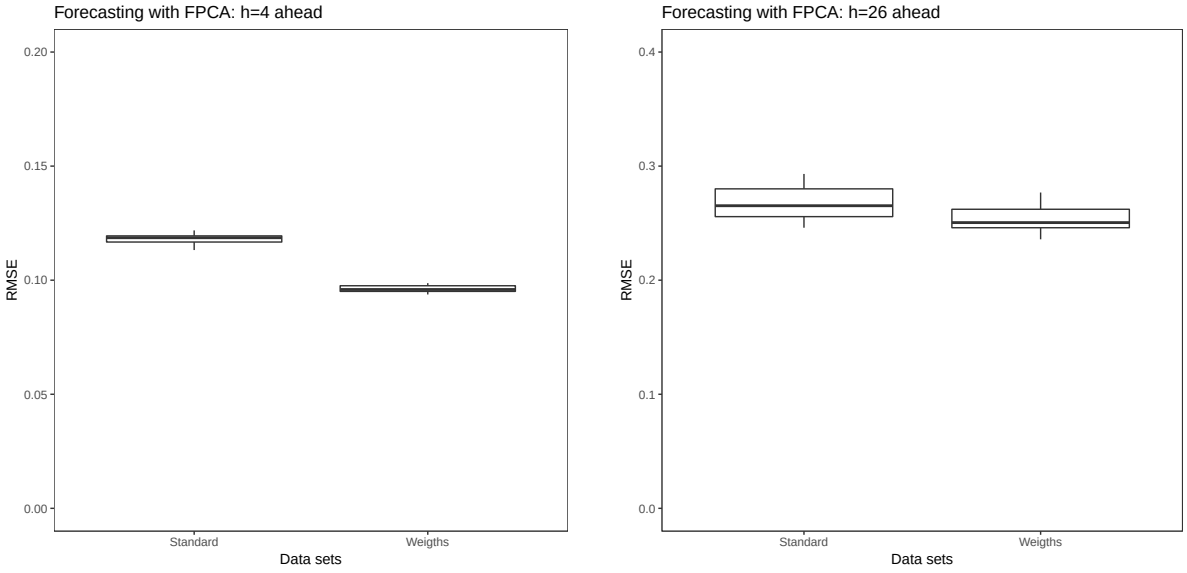


Figure 5.15: Boxplots of RMSE using the FPCA model for multiple step ahead forecasts of US Treasury data. Forecasting is conducted without weights (Standard) respectively with weights (Weights).

5.5 Gaussian Process prior model

Sambasivan and Das suggest that Gaussian process regression performs better to forecast yield curve in the medium and long term regions regarding maturities compared to other models. However, while their paper focuses on analyzing and comparing the forecasting performance at certain maturities across models, this thesis takes a different perspective in analyzing the forecasting performance of entire yield curves and comparing these.

5.5.1 Forecasting

For forecasting yield curves the implementation of the Dynamic Gaussian Process published by the Chennai Mathematical Institute on GitHub was used (Chennai Mathematical Institute, 2017). It is implemented in Python (Rossum, 1995) and based on the Gaussian processes framework `GPy` in Python (GPy, 2012). For the purpose of using the implementation within the applied study design of this thesis the code had to be adjusted (included in the electronic appendix.

Forecasting follows the two phases of the Dynamic Gaussian Process Algorithm using the GP model for regression. In the first phase at time step $t = 0$ the hyperparameters of the covariance function of the Gaussian Process $\mathbf{y}_0$ are estimated. This is carried out by maximizing the marginal log-likelihood of the model. Then the yield values for time step $t = 1$ are estimated based on Bayes theorem. The posterior covariance function and the posterior mean function are updated using the estimated hyperparameters and the estimated yield values. In the second phase for time steps $t \geq 1$ hyperparameters are estimated based on the updated process. The model is optimized by maximizing the marginal log-likelihood and estimates the yield values for time step t+1. Again covariance function and posterior mean function are updated forming the basis for the estimation of hyperparameters at the following step. With this iterating approach, starting with the second curve, every yield curve is forecasted.

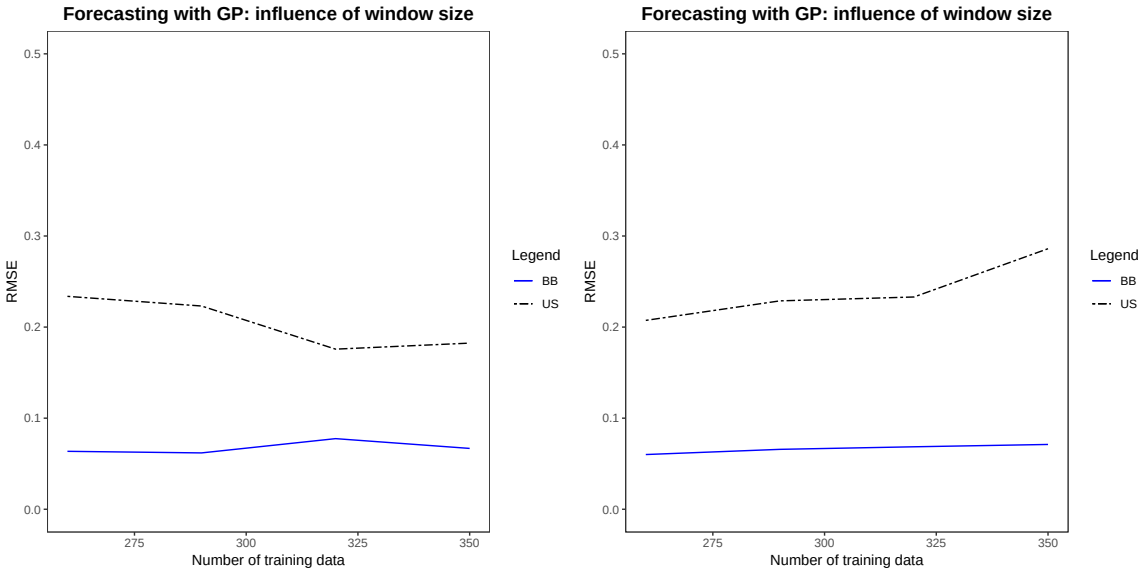
For modeling the covariance function a squared exponential kernel is used by the function `GPy.kern.RBF`, which is parameterized by parameters for length-scale and variance. In order to avoid during optimization these parameters becoming negative, all parameters for the GP model are constrained to be positive, which also includes the noise parameter.

Sambasivan and Das made *1-step* ahead predictions applying the Dynamic Gaussian Process. In order to extend the model to *h-step* ahead predictions, at every iterating step of the algorithm a loop is integrated, predicting yield values for h periods ahead based on the in-sample data and the generated predictions within this loop.

5.5.2 Results

First, the influence of the length of the training sample is examined. For the dynamic GP model using rolling windows of different length leads to variation in the mean RMSE, more for the short forecast horizon of $h = 4$ than the long horizon.

Evaluating the forecasting performance of the BB and US data it shows that the approach performs better for the BB than the US data, also displaying less variation.



(a) Forecast horizon 4.

(b) Forecast horizon 26.

Figure 5.16: Influence of window size on RMSE using the GP model for multiple step ahead forecasts of *Bundesbank* (BB) and US Treasury (US) data.

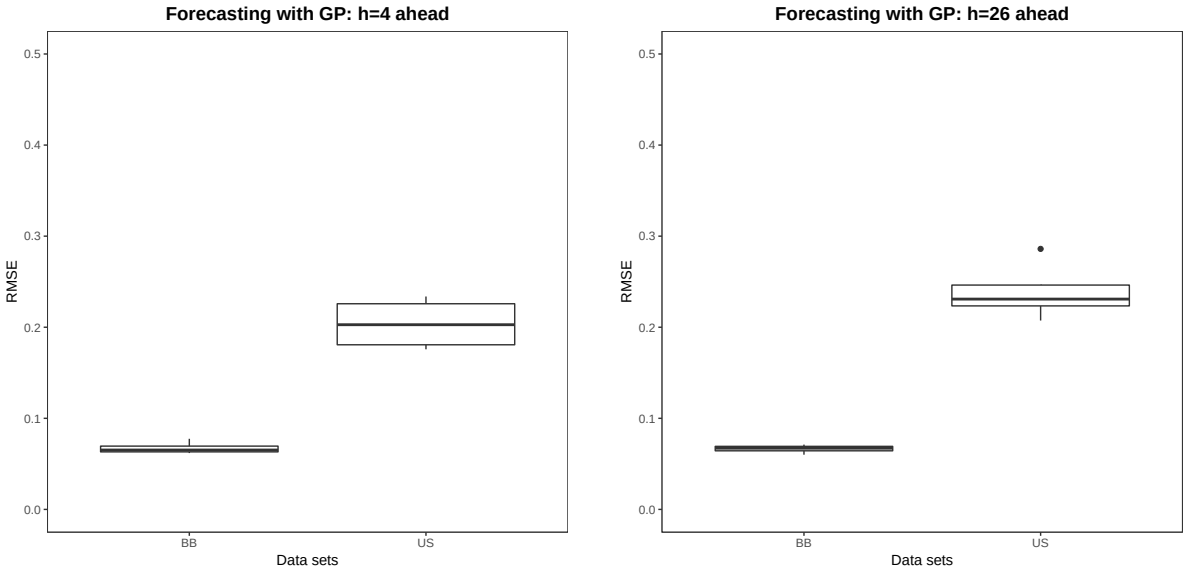


Figure 5.17: Boxplots of RMSE using the GP model for multiple step ahead forecast. Forecasts were conducted with 250 to 350 training periods for $h = 4$ and $h = 26$.

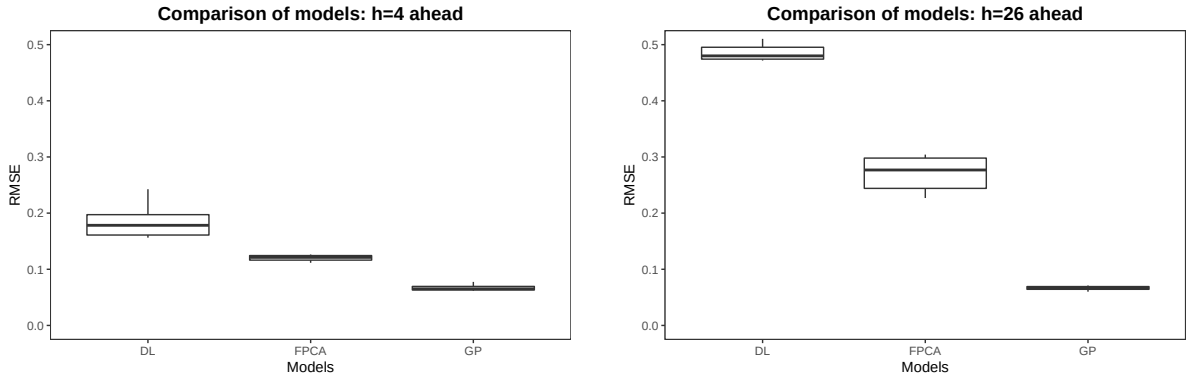


Figure 5.18: Boxplots of RMSE comparing different models for multiple step ahead forecasting of *Bundesbank* data.

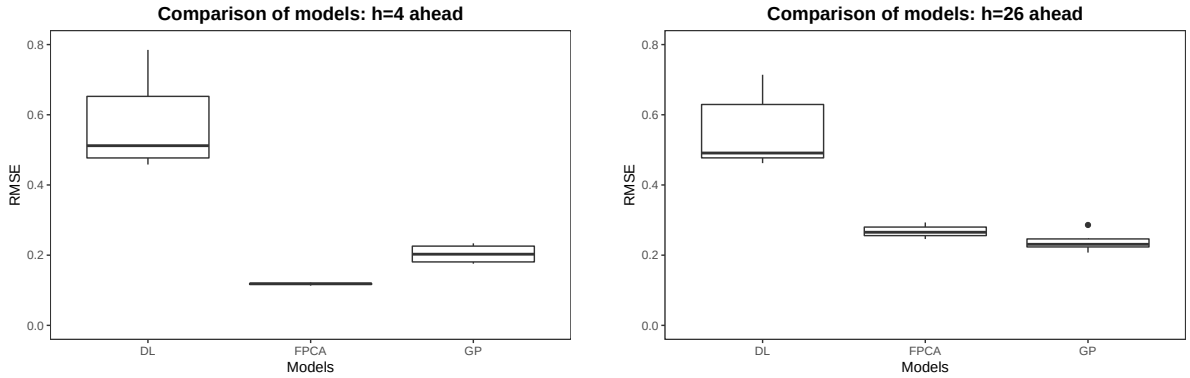


Figure 5.19: Boxplots of RMSE comparing different models for multiple step ahead forecasting of US Treasury data.

5.6 Comparison of results

When using rolling windows with a varying length l in a mid range of the available data the forecast performance is relatively constant. Using half of the available data produces good forecasting results.

Applying the DL model on different data sets as with BB and US data produces differing results, while the performance of FPCA does not vary much when applied to different data sets. From this follows that the FPCA model is better suited for a variety of data sets, varying in shape and also number of support points, than the DL model. The FPCA model profits from the smoothing of the curve before estimating the principal components.

For comparison in figures 5.18 and 5.19 the DL model with optimized AR(p), the standard FPCA and the dynamic GP model are displayed. The dynamic GP model performs best for the data set of the BB, remarkably better in the $h = 26$ forecasts. In case of forecasting US yield curves, this does not hold while the FPCA model performs comparably well respectively better. The FPCA model analyzes the more complicated US data very well to derive the scores subjected to forecasting.

Notably, the forecasted yield curves that are generated by the dynamic GP forecasting model show hardly any variation. The model, drawing on the last estimated mean function of the

training sample does not generate dynamic forecasts but relatively constant yield curves across the forecast horizon. As an example, figure 5.20 illustrates a $h = 26$ forecast by the FPCA model in comparison with the respective forecast of the dynamic GP model. While the dynamic GP models makes a smaller error in the short term regions of maturities, the FPCA performs better in the long term regions. However, analzing forecasting performance across the terms of maturities was not in the scope of this thesis.

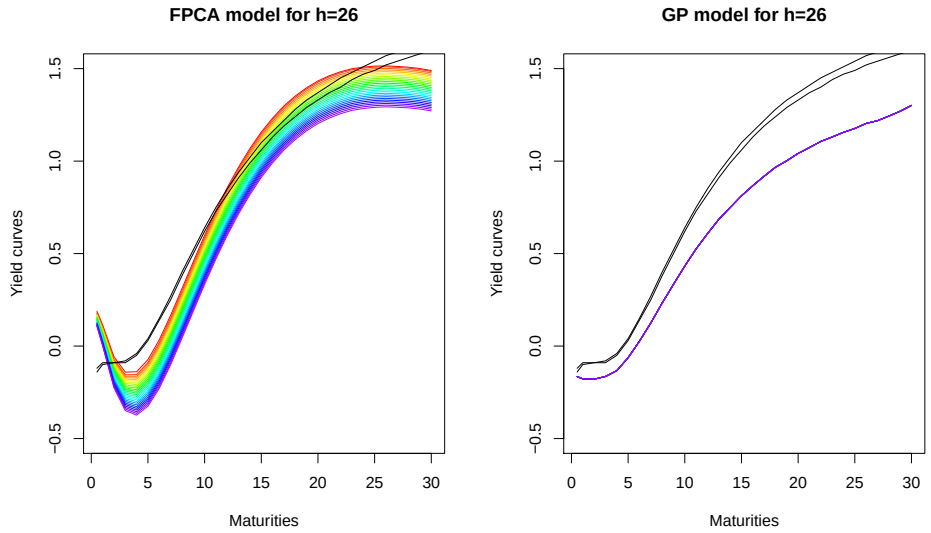


Figure 5.20: Variation in forecasts illustrated with training sample of 300 of *Bundesbank* data, last two observed yield curves in black.

6 Discussion and conclusion

This thesis examined three different models for forecasting yield curves and compared their forecasting performance by using two different data sets. The focus were methods that had an at least basic implementations in R and Python.
The DL model is a parametric model developed for the singular application on yield curves. As such, it has been widely accepted. However, efforts have been undertaken to improve the proposed framework as did Koopman et al. (2010). In this thesis the standard approach of forecasting the factors with an AR(1) model was extended to the application of an AR(p) model which generated notably better results, particularly for the *Bundesbank* data.
The approach of estimating functional principal components and forecasting the functional principal component scores as exercised by the FPCA model proposed in this thesis is applicable to a variety of functional data set and is not restricted to the case of yield curves. For the used data the model derived with the principal components the main modes of variation of yield curves as shown similarly for both data sets. In this study the model deliver constantly good results across different data sets and forecast horizon. A valuable contribution to forecasting performance using weights for more recent data could not be found in the framework of this study. It is also worth noting, that with the `"ftsa"` R package by Hyndman and Shang a well documented implementation for FPCA models for functional time series data exists.
The use of the dynamic GP model is regarded as rather experimental, because the referenced paper by Sambasivan and Das has not been regularly published but is published as a preprint. Particularly, it is developed for 1-step ahead estimation. Adapting the model to multi step ahead forecasting leads to notably good results, although the adapted model does not produce considerable variation within the forecasted curves. However, this persistent approach in the majority generates less errors when forecasting yield curves than the other models.
Out of the scope of this thesis was how the methods perform with respect to the different regions of terms of maturity. Additionally to the examined performance considering the entire yield curve, this would be another interesting subject of analysis. Also, the question of the required minimum window length of a training sample to obtain reliable forecasts could be explored.

References

Arbia, G. and Di Marcantonio, M. (2015). Forecasting interest rates using geostatistical techniques. *Econometrics*, 3(4):733–760.

Ažman, K. and Kocijan, J. (2005). Comprising prior knowledge in dynamic gaussian process models. *Proceedings of the International Conference on Computer Systems and Technologies (CompSysTech), Pages IIIB.2–1—IIIB.2–6*.

Benko, M., Härdle, W., and Kneip, A. (2009). Common functional principal components. *The Annals of Statistics*, 37(1):1–34.

Bergmeir, C., Hyndman, R. J., and Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, 120:70–83.

Bowsher, C. G. and Meeks, R. (2008). The dynamics of economic functions: Modeling and forecasting the yield curve. *Journal of the American Statistical Association*, 103(484):1419–1437.

Chen, Y. and Niu, L. (2014). Adaptive dynamic nelson–siegel term structure model with applications. 180(1):98–115.

Chennai Mathematical Institute (2017). Yield curve modeling using dynamic gaussian processes. GitHub repository. `https://github.com/cmimlg/YieldCurveModeling`.

Deutsche Bundesbank (1997). Schätzung von Zinsstrukturkurven. *Monatsbericht Oktober 1997*, pages 61–66.

Diebold, F. X. and Li, C. (2006). Forecasting the term structure of government bond yields. *Journal of Econometrics*, 130(2):337–364.

Diebold, F. X. and Rudebusch, G. D. (2013). *Yield curve modeling and forecasting: The dynamic Nelson-Siegel approach*. The Econometric and Tinbergen Institutes lectures. Princeton University Press, Princeton.

GPy (since 2012). GPy: A gaussian process framework in python. `http://github.com/SheffieldML/GPy`.

Guirreri, S. (2015). *YieldCurve: Modelling and estimation of the yield curve*. R package version 4.1. `https://CRAN.r-project.org/package=YieldCurve`.

Hall, P., Müller, H.-G., and Wang, J.-L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *The Annals of Statistics*, 34(3):1493–1517.

Hays, S., Shen, H., and Huang, J. Z. (2012). Functional dynamic factor models with application to yield curve forecasting. *The Annals of Applied Statistics*, 6(3):870–894.

Hyndman, R. J. and Shang, H. L. (2009). Forecasting functional time series. *Journal of the Korean Statistical Society*, 38(3):199–211.

Hyndman, R. J. and Shang, H. L. (2018). *ftsa: Functional Time Series Analysis*. R package version 5.2. `https://CRAN.R-project.org/package=ftsa`.

Klüppelberg and Sen (2010). Time series of functional data.

Koopman, S. J., Mallee, M. I. P., and van der Wel, M. (2010). Analyzing the term structure of interest rates using the dynamic nelson–siegel model with time-varying parameters. *Journal of Business & Economic Statistics*, 28(3):329–343.

Molenaars, T. K., Reinerink, N. H., and Hemminga, M. A. (2015). Forecasting the yield curve: art or science? *Magazine De Actuaris (The Actuary)*, 22(4):38–40.

Müller, H.-G., Stadtmüller, U., and Yao, F. (2006). Functional variance processes. *Journal of the American Statistical Association*, 101(475):1007–1018.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: `http://www.R-project.org/`.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer Series in Statistics. Springer Science+Business Media Inc, New York, NY, second edition edition.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian processes for machine learning*. Adaptive computation and machine learning. The MIT Press, Cambridge, Massachusetts and London, England.

Rossum, G. (1995). Python reference manual. Technical report, Amsterdam, The Netherlands.

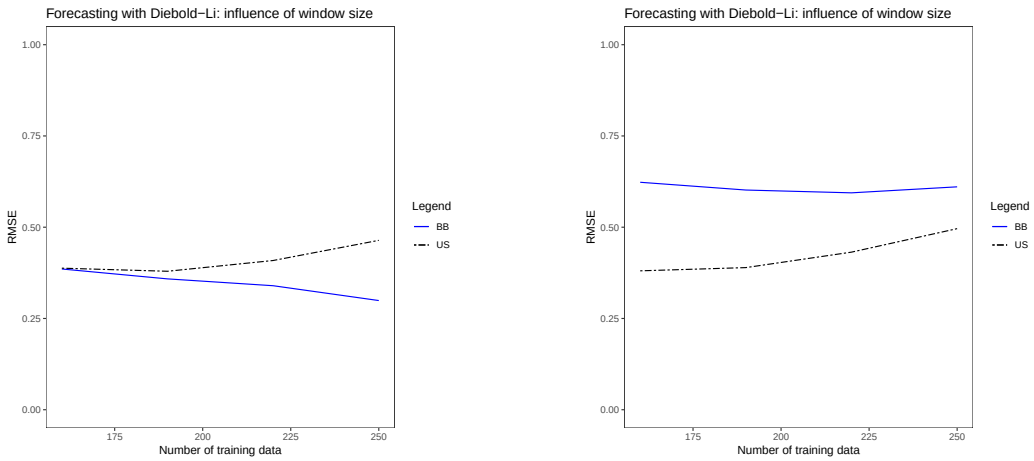
Ruppert, D. and Matteson, D. S. (2015). *Statistics and Data Analysis for Financial Engineering - with R examples*. Springer, Berlin, Heidelberg.

Sambasivan, R. and Das, S. (2017). A statistical machine learning approach to yield curve forecasting. arXiv preprint arXiv:1703.01536.

Shang, H. L. (2013). ftsa: An R package for analyzing functional time series. *The R Journal*, 5(1):64–72.

Williams, C. K. I. (1997). Prediction with gaussian processes: From linear regression to linear prediction and beyond. In *Learning and Inference in Graphical Models*, pages 599–621. Kluwer.

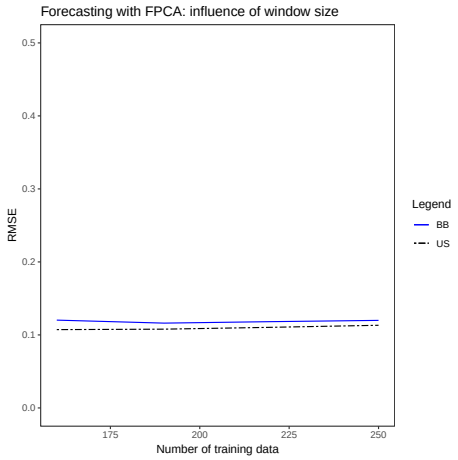
Appendix



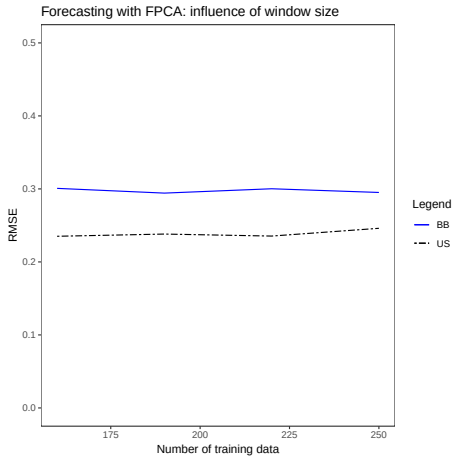
(a) Forecast horizon 4.

(b) Forecast horizon 26.

Figure 6.1: Influence of smaller window size on RMSE using the DL model for multiple step ahead forecasts of *Bundesbank* (BB) and US Treasury (US) data.

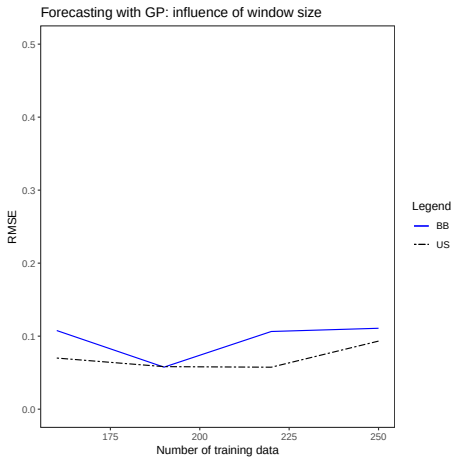


(a) Forecast horizon 4.

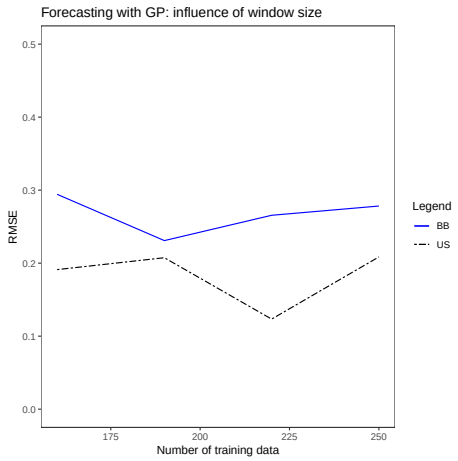


(b) Forecast horizon 26.

Figure 6.2: Influence of smaller window size on RMSE using the FPCA model for multiple step ahead forecasts of *Bundesbank* (BB) and US Treasury (US) data.



(a) Forecast horizon 4.



(b) Forecast horizon 26.

Figure 6.3: Influence of smaller window size on RMSE using the GP model for multiple step ahead forecasts of *Bundesbank* (BB) and US Treasury (US) data.

Electronic appendix

The electronic appendix comprises all the code and data sets to reproduce the results referenced in this thesis.
Code and data are stored in the folders *R_MA_Reinicke* and *Py_MA_Reinicke* comprising files for R code and for Python code, respectively.

R_MA_Reinicke
In `03_function_compare` the working directory has to be set. Running the file also loads `01_packages` and `02_input_data`. The file includes the functions to generate the forecast models DL and FPCA and functions associated with the rolling window study design. With `04_empirical_analyses` different modeling set-ups can be computed. (For variation of FPCA model "weight = TRUE" has to be set in file `03_function_compare`). `05_plots` includes descriptive plots and informative plots for the DL and FPCA model. In `06_plots_analyses` the plots for forecast results are found. "results_plots" contains the results used.
`07_function_forecasting_variation` shows plots of the variation within forecasts; Associated with this file are the csv-files "BB_GPpred_estimates_4_(300)", and "BB_GPpred_estimates_26_(300)". `08_footnotes_appendix` contains various calculations and plots. "results_plots_appendix" contains the results used.
`02_input_data` loads data sets "BB_2009_2018" and "US_2009_2018" and makes necessary adaptions. It also exports the data to files for the Python code; this is unabled, because the data is provided in the respectiv folder.

Py_MA_Reinicke
With `01_GP_input_data` the data is loaded. `02_GP_function_rolling` includes the function `roll_pred` required for the rolling study with the dynamic GP model. *03_BB_Data_GP_predh* works independently and comprises function `h_pred_BB` for h-step ahead forecasts for BB data. In `04_GP_analyses` different modeling set-ups can be computed with varying length of training sample windows. The data files in csv format are generated by the R file `02_input_data`. For loading, the correct working directory has to be set.

Statutory Declaration

I herewith declare that I have composed the present master thesis myself and without use of any other than the cited sources and aids. Sentences or parts of sentences quoted literally are marked as such; other references with regard to the statement and scope are indicated by full details of the publications concerned. The master thesis in the same or similar form has not been submitted to any examination body and has not been published. This master thesis was not yet, even in part, used in another examination or as a course performance.

München, April 16, 2019 _____